UNIVERSIDADE TÉCNICA DE LISBOA

INSTITUTO SUPERIOR DE ECONOMIA E GESTÃO

MESTRADO EM ECONOMETRIA APLICADA E PREVISÃO

Impacto do *heaping* em modelos para dados de duração

Félix Ferreira Bernardo

Orientação: Doutor José Manuel de Matos Passos

Júri:

Presidente: Doutor José Manuel de Matos Passos, professor auxiliar do Instituto Superior de Economia e Gestão da Universidade Técnica de Lisboa;

Vogais: Doutora Esmeralda de Jesus Ratinho Lopes Arranhado Ramalho, professora auxiliar da Universidade de Évora;
Doutor Montezuma Boaventura Guimarães Dumangane, professor auxiliar do Instituto Superior de Economia e Gestão da Universidade Técnica de Lisboa;

Março / 2007

UNIVERSIDADE TÉCNICA DE LISBOA

INSTITUTO SUPERIOR DE ECONOMIA E GESTÃO

MESTRADO EM ECONOMETRIA APLICADA E PREVISÃO

Impacto do *heaping* em modelos para dados de duração

Félix Ferreira Bernardo

Orientação: Doutor José Manuel de Matos Passos

Júri:

Presidente: Doutor José Manuel de Matos Passos, professor auxiliar do Instituto Superior de Economia e Gestão da Universidade Técnica de Lisboa;

Vogais: Doutora Esmeralda de Jesus Ratinho Lopes Arranhado Ramalho, professora auxiliar da Universidade de Évora;
Doutor Montezuma Boaventura Guimarães Dumangane, professor auxiliar do Instituto Superior de Economia e Gestão da Universidade Técnica de Lisboa;

Março / 2007

Resumo

Dados de duração são muitas vezes obtidos recorrendo a questionários em que os sujeitos são inquiridos sobre aspectos do seu passado. Dados retrospectivos podem conter vários tipos de imprecisões uma delas é o aparecimento de frequências anormalmente elevadas para certas durações. Na literatura este tipo de anormalidade é conhecido por *heaping* . Alguns trabalhos anteriores demonstraram que até *heaping* simétrico pode conduzir à inconsistência dos estimadores de máxima verosimilhança. No trabalho que se segue aprofundamos, recorrendo a simulação de Monte Carlo, as consequências do *heaping* na estimação dos parametros de um modelo Weibull para dados de duração completos e censurados à direita. Comparamos ainda as consequências, para a estimação por máxima verosimilhança, de diferentes percentagens de *heaping* e diferentes dimensões de censura.

Palavras chave: *Heaping* , Modelos de duração, Hazard proporcional, Weibull, Censura à direita, Simulação de Monte Carlo.

Classificação JEL: C15, C41.

Abstract

Duration data are, many times, acquired by questionnaires where the subjects are inquired about some aspects of its past. Retrospective data may contain some form of response error where the most notorious are the appearance of abnormally high frequencies for certain durations. In the literature this type of abnormality is known as heaping. Some previous works have shown that even symmetrical heaping can lead to inconsistency of maximum likelihood estimators. In the work that follows we use Monte Carlo simulation to study the consequences of heaping in the estimation of the parameters of a Weibull model with completed and right censored duration data. We evaluate also the consequences, for maximum likelihood estimation, of different percentages of heaping and different dimensions of censoring.

Keywords: Heaping, Duration Models, Proportional Hazard, Weibull, Right Censoring, Monte Carlo simulation.

JEL Classification: C15, C41.

Aos Professores José Passos pela orientação e Montezuma Dumangane pela sugestão do tema da tese, aos colegas Mónica e Tiago pelo incentivo e amizade e finalmente à minha família pela paciência.

Conteúdo

Capítulo 1

Introdução

Dados de duração são muitas vezes obtidos recorrendo a inquéritos em que os sujeitos são inquiridos sobre aspectos do seu *passado*. Contudo dados retrospectivos contêm muitas vezes vários tipos de imprecisões (dados em falta, variáveis medidas com erro, frequências anormalmente altas para algumas observações, etc...).

É já abundante a literatura sobre a má qualidade dos dados sobre desemprego, a título de exemplo, sobre o GOSEP (German Socio-Economic Panel) [1], ver Wolff a Augustin (2003), em dados antropométricos de crianças tanzanianas , Heitjan e Rubin (1990) encontram demasiadas idades múltiplas de 6 meses, sobre consumo das familias italianas , Battistin, Miniaci e Weber (2003) encontram frequências anormais para consumos múltiplos de 100.000 *Lit*, entre outros.

Os inquiridos podem não reportar acontecimentos simplesmente por esquecimento ou porque não lhes atribuem importância. Respondentes podem também, consciente ou inconscientemente, redefinir o passado ajustando-o por exemplo às suas "preferências". Por exemplo, nos inquéritos sobre desemprego, muitas mulheres redefinem o seu estado como "donas-de-casa" quando na verdade estiveram desempregadas (Jürges, 2004).

De acordo com a psicologia cognitiva a precisão com que se recorda acontecimentos passados depende principalmente de três factores: interferência (ocorrência de acon-

[1] Frequências anormais, de entradas no estado de desemprego em Janeiro e de saídas em Dezembro.

tecimentos similares), intervalo de tempo decorrido desde o acontecimento e saliência do acontecimento, i. é, a importância do acontecimento para o sujeito [2]. Para um estudo mais aprofundado sobre a importância da saliência ver Jürges (2004).
No âmbito deste trabalho interessa sobretudo reter que os erros na descrição de acontecimentos passados *dependem da duração* dos acontecimentos a reportar.

1.1 Heaping

1.1.1 Definição

Incomplet data (Little & Rubin, 2002) é um vasto tema de estudo em Econometria em particular o campo de estudos sobre *missing data*, Rubin (1976), em que os dados são parcial ou totalmente desconhecidos; para exemplos e procedimentos como lidar com este tipo de dados ver Ramalho e Smith (2003)
Contudo há muitas situações ambíguas, i. é, os dados nem são completamente conhecidos nem são completamente desconhecidos; em vez de todo o espaço de resultados possíveis apenas se observa uma parte desse espaço. Clarificando: se como habitualmente estivermos interessados na distribuição de uma variável Y, condicional no vector de variáveis $X = [X_1, ..., X_n]$, observamos (n+1)-uplos $(y_i, x_{i1}, ..., x_{in})$ que deveriam poder assumir qualquer valor no espaço de resultados de $[Y, X]$ mas na verdade por limitações óbvias (sobretudo no caso de variáveis contínuas) é apenas observado um subconjunto desse espaço. Se por exemplo Y_i "mede"a altura do indivíduo y_i dificilmente poderá assumir o valor $\frac{\pi}{2}m$ embora seja tão provável como $1,57m$. Heitjan e Rubin (1991) referem-se a este tipo de dados como *coarse data*.
Coarse data podem ser gerados de várias formas sendo a mais vulgar o arredondamento, neste caso observa-se apenas um subconjunto contável [3] do espaço de resultados de $[Y, X]$. Note-se que na prática temos quase sempre dados deste tipo uma vez que

[2] Para mais detalhe ver Gleitman (1993).
[3] Finito ou infinito numerável

mesmo aumentando a precisão da descrição de uma variável contínua teremos de parar nalguma casa decimal. De notar ainda que dados censurados e dados agrupados são também casos particulares de *coarse data*. De facto é uma questão de escala, considera-se que estamos perante *coarse data* se a escala de *coarseness* observada é mais grosseira do que aquela que foi definida pelo observador.

Na literatura é muitas vezes referido um tipo de *coarse data* conhecido por *heaping data* , *dados amontoados* , onde se inclui em particular o caso da *preferência* de digito (muitas vezes ditada por razões culturais).

Para Heitjan e Rubin (1991), **um conjunto de dados diz-se *heapead*, amontoado, se incluir items reportados com vários tipos de *coarseness***. Contudo esta definição é pouco operativa embora no texto citado ela praticamente não seja utilizada. Os autores estão sobretudo interessados em derivar um resultado para *coarse data*, a propriedade CAR (*coarseness at random*) , semelhante à propriedade MAR (*Missing at random*) para *missing data*, com o objectivo de tratar *coarse data* como equivalentes a dados agrupados e utilizar o quadro teórico correspondente.

Trabalhos subsequentes tendem a definir um conjunto de dados como *heapead* se empiricamente se verifica **um anormal** [4] **amontoado de dados junto a determinados pontos do espaço de resultados** (se os dados são períodos de duração) **ou junto a certas datas** (se for utilizada uma metodologia de eventos num calendário). São particularmente ilustrativos, os trabalhos de Torelli e Trivellato (1993), onde os autores modelam dados retirados de LFS's (Labor Force Surveys) da Lombardia e observam frequências anormalmente elevadas nas durações múltiplas de 6 meses ou o trabalho de Krauss e Steiner (1998) sobre o GOSEP onde os autores observam saídas anormalmente elevadas do estado de desemprego em Dezembro e entradas anormalmente elevadas em Janeiro.

[4] Anormal em relação ao esperado à priori ou esperado a partir de dados de outra fonte

1.1.2 Evidência e consequências

A partir de dados empíricos nem sempre é fácil inferir o padrão de *heaping* , em particular, torna-se por vezes difícil separar os efeitos sazonais e/ou outros efeitos de contexto (como por exemplo nos dados de desemprego, do natural aumento da taxa de transição no final do período de benefício de prestação de desemprego) do *heaping* propriamente dito.

Estudando dados de desemprego obtidos do GOSEP (German Socio-Economic Panel) Jürgens (2004), Kraus e Steiner (1998) entre outros, encontram evidência de *heaping* . Os autores citados recorrem a informação externa aos dados, inscrições no FLO (Federal Labor Office) [5], para estimar a percentagem de observações *amontoadas*. Jürgens (2004), referindo-se a estudos anteriores observa que a percentagem de dados *amontoados* pode variar muito até corresponder praticamente à totalidade dos dados. Torelli e Trivellato (1993), estudam dados de desemprego de jovens da Lombardia italiana e verificam picos anormais para durações de desemprego múltiplas de 6 *meses*. Sendo comum, o *heaping* é em geral caracterizado por um padrão. Caracterizar esse padrão fica fora do âmbito deste trabalho, diremos apenas que em geral os dados o sugerem com clareza (múltiplos de 6 para escalas mensais, unidades em vez de números decimais, etc...). Contudo, apesar do padrão de *heaping* ser, em geral, um assunto em aberto, há abundantes evidências empíricas de que *heaping* é a regra e não a excepção. Mas quais os efeitos do *heaping* nos modelos para dados de duração?

Tentar responder a esta pergunta é em parte o propósito deste trabalho.

Intuitivamente, um efeito óbvio do *heaping* é a redução da variância amostral e do desvio padrão (SE) dos estimadores, afinal, o amontoar de dados reduz a variabilidade dos mesmos. Por outro lado talvez alguns tipos de *heaping* não provoquem enviesamento nos estimadores dos parâmetros (pelo menos os que forem obtidos por métodos lineares) [6]. Regra geral, em modelos para dados de duração, os métodos de estimação são não lineares e portanto o enviesamento poderá ocorrer até no caso de

[5] Inscrição necessária para receber subsídio de desemprego.

[6] Por exemplo *heaping* simétrico em torno dos pontos de *heaping*

heaping simétrico. Acessoriamente, ao "reduzir"o tamanho da amostra, sobretudo em amostras pequenas, o *heaping* pode gerar problemas na estimação (não convergência dos algoritmos usados).

1.1.3 Estratégias

Várias estratégias têm sido adoptadas para lidar com o problema da falta de qualidade dos dados, e do *heaping* em particular, a mais popular sem dúvida é **ignorá-lo**.
Num trabalho sobre dados antropométricos de crianças tanzanianas, Heitjan e Rubin (1990), observam que a idade das crianças é amontoada nos múltiplos de 6 meses e propõem uma abordagem bayesiana em que a *prior* é obtida a partir dos dados e a distribuição à posteriori é obtida por múltipla imputação de dados gerados pelo modelo.
No trabalho já referido para *coarse data* em geral, Heitjan e Rubin (1991), estabelecem condições suficientes para que dados de *heaping* [7] possam ser tratados como dados agrupados e os resultados de inferência (bayesiana) permaneçam válidos. Essencialmente deve observar-se a propriedade CAR (coarseness at random), o que parece pouco plausível uma vez que em geral o *heaping* depende pelo menos da duração.
Numa abordagem teórica do fenómeno de *digit preference* Ryu e Slotje (2000) usam o método de máxima entropia com as condições dos momentos impostas a partir dos momentos amostrais. Admitindo que os momentos amostrais de menor ordem são pouco afectados pela preferência de digito "alisam"a distribuição das durações cujo logaritmo pode ser escrito em série de potências [8]. Note-se que a utilização de um kernel de alisamento não serve o mesmo propósito uma vez que apesar de suavizar a distribuição, no lugar dos picos ficam máximos locais.
Outra abordagem possível para o problema do *heaping* é tratá-lo como um caso de variável medida com erro e usar as ferramentas desenvolvidas para este tipo de

[7] Note-se que *heaping data* é um caso particular de *coarse data*.

[8] $ln(f(t)) = \sum_{i=0}^{I} a_i t^i$

modelos, como exemplo (Augustin & Schneeweiß, 2005).
Torelli e Trivellato (1993), Kraus e Steiner (1998), Augustin e Wolff (2003) e Gill, Petoussis e Zeelenberg (2004) assumem um modelo de variável latente não observável,

$$T_i^* = T_i + \delta_i Y_i$$

onde T_i^* é a duração observada, T_i a verdadeira duração , Y_i é uma variável de Bernoulli que toma o valor 1 se a observação i é *amontoada* e $\delta_i = h_e - T_i$ onde h_e é o ponto do conjunto de *heaping* mais próximo de T_i. Como vamos adoptar esta abordagem, o trabalho destes autores será exposto no próximo capítulo.

1.2 Impacto do Heaping em modelos para dados de duração

Neste trabalho procuraremos estudar, por simulação de Monte Carlo, o impacto do Heaping em modelos para dados de duração, conferindo e se possível extendendo os resultados obtidos por outros autores. No essencial usaremos o modelo proposto por Torelli e Trivellato (1993).
Esta trabalho está estruturado da seguinte forma: no próximo capítulo, "**Modelo teórico**", resumiremos alguns trabalhos anteriores relevantes e faremos o enquadramento teórico do problema, no capítulo seguinte, "**Estudo de simulação**", generalizaremos alguns dos resultados obtidos por outros autores, alargando o conjunto de variáveis explicativas a três casos comuns, variáveis simétricas, variáveis com assimetria positiva e variáveis dummy, finalmente no último capítulo, "**Conclusões**", apresentaremos as conclusões.
Sempre que possível, os quadros de resultados e figuras, serão apresentados e comentados aquando da sua primeira referência.
Para não sobrecarregar a exposição alguns resultados serão apresentados em anexo.

Capítulo 2

Modelo teórico

2.1 Trabalhos anteriores

2.1.1 Escolha do conjunto de *heaping*

De acordo com Torelli e Trivellato (1993), endogeneizar o conjunto de pontos de *heaping* , H, não parece ser uma alternativa viável; citando os autores, "*...não é simples endogeneizar H . Patentemente, aparecem problemas de identificabilidade. As distribuições empíricas das durações reflectem a convolução de três processos estocásticos: um processo que governa as taxas de entrada no estado, um processo que governa o período de permanência no estado, e um processo segundo o qual cada duração é reportada. Mesmo assumindo que a taxa de entrada é constante e as durações seguem um caminho conhecido (no sentido de forma funcional, mesmo que com parâmetros desconhecidos), uma investigação genérica das condições de identificabilidade para o processo de heaping e para os parâmetros desconhecidos do processo das durações é árduo,...*".

A escolha do conjunto de *heaping* é essencialmente um processo empírico.

Admitamos por isso que os dados sugerem amontoados "anormais"de durações pertencentes ao conjunto $S = \{h_1, h_2, ..., h_k\}$, pretendendo testar essa escolha para o conjunto de valores de *heaping* , Brewer e Roberts (2001) propõem duas estatísticas e respectiva metodologia de teste. Com efeito, considerando $t_1, t_2, ..., t_n$ as durações observadas, com $1 \leq t_i \leq M$ e $S = \{h_1, h_2, ..., h_k\}$, uma escolha para conjunto de *heaping* . Sejam , $f(i)$ o número de respostas i e

$$R = \{A \in \mathscr{P}(\{1, 2, ..., M\}) : \ \sharp A = k\} \ ,$$

defina-se [1]

$$C_1(i) = \begin{cases} f(i) - \dfrac{f(i+1) + f(i-1)}{2} & se \quad i \neq 1, M \\ f(1) - f(2) & se \quad i = 1 \\ f(M) - f(M-1) & se \quad i = M \end{cases}$$

$C_2(1) = 1$ se $f(1) \geq f(2)$

$$C_2(i) = \begin{cases} 1 \quad se \quad f(i-1) \leq f(i) \geq f(i+1) \\ 0 \qquad\qquad\qquad cc \end{cases} \qquad para\ i \in \{2, 3, ..., M-1\}$$

e $C_2(M) = 1$ se $f(M) \geq f(M-1)$

Considere-se $C_j(A) = \sum_{i \in A} C_j(i) \quad para\ j = 1, 2$

e as estatísticas

$$H_j(S) = \frac{C_j(S) - E_R\{C_j(R)\}}{max_R\{C_j(R)\} - E_R\{C_j(R)\}} \quad para\ j = 1, 2$$

Definidas desta forma $H_j(S) \leq 1$. Valores elevados de $H_j(S)$ indicam evidência de *heaping* para valores de S, $H_j(S) = 0$ indica que S apresenta o valor esperado de *heaping* de um qualquer subconjunto de k elementos enquanto valores negativos

[1] Os autores, por uma questão de optimização restringem este conjunto uma vez que é pouco plausível que dois valores consecutivos, i e $i+1$ sejam ambos valores de *heaping* e por isso elementos que não verifiquem esta restrição não foram incluídos em R

indicam menos evidência de *heaping* para valores de S.

Na hipótese nula do *heaping* seguir um padrão aleatório vs. a hipótese alternativa de *heaping* em S toma-se como valor-p a proporção de subconjuntos R tais que $C_j(R) \geq C_j(S) \quad j = 1, 2$, i.é a proporção de casos mais desfavoráveis para H_0.

2.1.2 Torelli e Trivellato

Torelli e Trivellato num trabalho de 1993, "Modelling inaccuracies in job-search duration data", procuram pela primeira vez modelar o *heaping* presente em dados de duração provenientes de inquéritos ao emprego de jovens da Lombardia. No artigo citado, Torelli e Trivellato propuseram um modelo de variável latente para acomodar o *heaping* . Na sua formulação mais simples:

Dada uma variável não negativa T que descreve as durações, com uma fdp $f(t, \theta)$ conhecida a menos de um vector de parâmetros θ. Se t_i forem as verdadeiras durações, t_i^* as durações observadas e $H = \{h_1,\ h_2,\ \ldots\ , h_M\}$ o conjunto de *heaping* , ordenado de forma crescente, os autores propõem o modelo:

$$t_i^* = t_i + \delta_i . y_i \tag{2.1}$$

com Y_i uma variável de Bernoulli que toma o valor 1 se a observação i é *amontoada* e $\delta_i = h_e - t_i$ onde h_e é o ponto do conjunto de *heaping* mais próximo de t_i.

O parâmetro de Bernoulli associado à v. a. Y_i é modelado por $P(Y_i = 1) = G(t_i, \gamma)$, onde G, a função de *heaping* , é conhecida a menos de um vector de parâmetros γ.

Recorrendo a um estudo de simulação [2] os autores estudam a consistência de três estimadores de máxima verosimilhança, A, B e C:

- A, recorre à verosimilhança derivada de (2.1);
- B, usa uma função de *heaping* $G(t, \gamma)$ constante por troços (o que permite factorizar a verosimilhança e simplifica a estimação);
- C, ignora o *heaping* .

[2] Usaram 100 réplicas e que na nossa reprodução do trabalho destes autores se revelou claramente insuficiente!

Para T escolhem três distribuições, exponencial, Weibull e log-logística com $\theta = [\beta_0]$; para G escolhem a exponencial com $\gamma = [\gamma_0]$ e para H, num estudo H_1 múltiplos de 12 (noutro estudo H_2 múltiplos de 6).
Fazem variar γ e θ e o conjunto de *heaping* H; o que conduz em geral a mais de **50% de observações *amontoadas***, o que parecendo adequado aos dados empíricos que os autores modelam como exemplo, nos parece, **em geral um (padrão de) *heaping* excessivo**, concluem:

- Se T seguir uma distribuição exponencial todos os estimadores são consistentes;

- Se T seguir uma distribuição Weibull ou log-logística B e C são inconsistentes, B por defeito e C por excesso;

- O enviesamento diminui quando se refina H.

- O enviesamento está positivamente correlacionado com a percentagem de observações *amontoadas*.

Finalmente, e na modelação dados reais, os autores propõem em alternativa a utilização de *dummy's*, uma para $t_i = 12k$ e outra para $t_i = 12k + 6$, $(k = 1, 2, ...)$ mas os parâmetros estimados revelam-se inverosímeis e desaconselham esta estratégia.

2.1.3 Kraus e Steiner

Kraus e Steiner (1998) adaptam o modelo da secção anterior para dados obtidos a partir de uma metodologia baseada em eventos num calendário. Neste tipo de inquéritos o padrão de *heaping* não se refere a frequências inesperadamente elevadas de certas durações mas sim a um amontoado inesperado de inícios e/ou fins de certos eventos em determinados pontos do calendário. No caso *Dezembro* para os finais e *Janeiro* para os inícios.
Estudando dados do GOSEP os autores modelam os inícios e os finais de durações de desemprego com dois conjuntos de *heaping* (um para os inícios e outro para os

finais) reduzidos a um ponto de *heaping* , a saber, $H_E = \{Dezembro\}$ para os finais e $H_B = \{Janeiro\}$ para o início dos eventos. Quanto ao padrão de *heaping* , os inícios em Fevereiro e Março são *amontoados* para Janeiro e os finais de Outubro e Novembro são *amontoados* para Dezembro.

Resumindo o modelo proposto, e assumindo os dados como discretos, consideraram as variáveis aleatórias,

T_B o verdadeiro mês de entrada no estado [3] e T_B^* o mês observado;

T_E o verdadeiro mês de saída do estado e T_E^* o mês observado.

Tal como para Torelli e Trivellato, os $\delta' s$ referem-se ao padrão de *heaping* , a saber, $\delta(T_B) = Janeiro - T_B$ e $\delta(T_E) = Dezembro - T_E$; $Y's$ são variáveis de Bernoulli que tomam o valor 1 se a observação for *amontoada* com os respectivos parâmetros governados pelas funções de *heaping* $G_B(t_B)$ e $G_E(t_E)$, respectivamente.

Se T for a verdadeira duração e T^* a duração observada

$$\begin{aligned} T_B^* &= T_B + \delta(T_B).Y_B(T_B) \\ T_E^* &= T_E + \delta(T_E).Y_E(T_E) \\ logo: & \\ T^* &= T + \delta(T_E).Y_E(T_E) - \delta(T_B).Y_B(T_B) \end{aligned} \tag{2.2}$$

Quanto às funções de *heaping* $G_B(t_B)$ e $G_E(t_E)$, os autores tentam estimar os seus parâmetros em conjunto com os restantes parâmetros do modelo mas devido a problemas numéricos (poucas observações para certos meses) acabam por estimá-las recorrendo a informação externa, registos no *Federal Labor Office (FLO)*, a saber:

Se $R_S(t_B)$ for a frequência relativa de uma entrada no estado de desemprego reportada no GOSEP e $R_O(t_B)$ for a frequência relativa de uma entrada no estado de desemprego reportada nos registos oficiais; se $R_S(t_E)$ for a frequência relativa de uma saída do estado de desemprego, reportada no GOSEP e $R_O(t_E)$ for a frequência relativa de uma saída do estado de desemprego, reportada nos registos oficiais, os autores definem os estimadores, para pontos que não sejam de *heaping* ,

$$\widehat{G}_B(t_B) = 1 - \frac{\widehat{R}_S(t_B)}{\widehat{R}_O(t_B)} \quad e \quad \widehat{G}_E(t_E) = 1 - \frac{\widehat{R}_S(t_E)}{\widehat{R}_O(t_E)} \ ,$$

[3] Desemprego

tendo obtido, estatisticamente não nulas, as estimativas :
Para os inícios, Fevereiro, 0.4394, Março, 0.3207 e para on finais, Outubro, 0.457 e Novembro, 0.5805 [4]. O que conduz a **um padrão de *heaping* bastante mais reduzido do que o estudado por Torelli e Trivellato** (1993) [5] e além disso **as durações aumentam no máximo 4 meses**, i. é, em geral menos do que no caso de Torelli e Trivellato.
Poder-se-ia perguntar se ainda vale a pena optar por acomodar o *heaping* no modelo.

Recorrendo à metodologia de hazard proporcional e à modelação semiparamétrica, $h(t) = h_0(t)exp(-X.\beta)$, para dados discretos e a algoritmos de maximização da verosimilhança, os autores estimam os parâmetros das variáveis explicativas mais comuns na modelação de dados de desemprego. Avaliam três modelos alternativos:
- $Modelo\ A$ que ignora o *heaping* ;
- $Modelo\ B$ que usa as funções de *heaping* estimadas;
- $Modelo\ C$ que usa funções de *heaping* arbitrárias.

Os coeficientes das variáveis explicativas (faixa etária, estado civil, grau de educação, *dummy's* de Janeiro e Dezembro, etc...) estimados pelos três modelos diferem muito pouco! Apenas as *dummy's* de Janeiro e Dezembro diferem de forma significativa, sobretudo em relação ao modelo C, i. é, parecem *modelar* o *heaping* no modelo que não o modela. De seguida avaliam até que ponto o efeito do *heaping* depende da especificação da *baseline hazard*. Como alternativa à modelação semiparamétrica usual os autores flexibilizam a forma funcional da *baseline hazard* e modelam-na como uma *logit* de uma fracção racional de polinómios da duração, tomam $h_0(t) = \frac{exp(t+t^2+\frac{1}{t})}{1+exp(t+t^2+\frac{1}{t})}$, tendo concluído que os coeficientes das variáveis explicativas não se alteram significativamente.
Finalmente, procurando avaliar a qualidade relativa dos modelos A e B e uma vez que o modelo A pode ser visto como um caso particular do modelo B com as funções de *heaping* iguais a zero, os autores implementaram um teste de rácio de verosimilhanças

[4] Além disso, por serem pontos de *heaping* , as estimativas de Dezembro, para os finais e Janeiro, para os inícios, são iguais a 1.

[5] Cerca de 11% dos inícios sofrem de *heaping* .

tendo concluído que o modelo que acomoda o *heaping* se ajusta melhor aos dados [6].

2.1.4 Augustin e Wolff

Num trabalho recente, Augustin e Wolff (2003), procuram enquadrar teoricamente os trabalhos de Torelli e Trivelhato (1993) (*heaping* enviesa parâmetros do modelo Weibull) e de Kraus e Steiner (1998), em particular a surpreendente insensibilidade ao *heaping* dos dados do GOSEP.

Estudando o modelo Weibull , de função densidade de probabilidade,

$$f(t_i) = \alpha\lambda t_i^{\alpha-1} exp(-\lambda t_i^{\alpha})$$

(de que o modelo exponencial é um caso particular quando $\alpha = 1$) com amostragem aleatória e o vector de variáveis explicativas reduzido à constante e $\lambda = exp(-X.\beta)$. Fixando α, o estimador da máxima verosimilhança de β pode ser obtido explicitamente:

$$\hat{\beta} = \frac{1}{\alpha} \ln(\frac{\sum_{i=1}^{n} T_i^{\alpha}}{n})$$

Mas no caso de *heaping* , as verdadeiras durações, T_i , não são observáveis, o que se observa é T_i^*.
Um estimador *naïve* pode ser:

$$\hat{\beta}^* = \frac{1}{\alpha} \ln(\frac{\sum_{i=1}^{n} (T_i^*)^{\alpha}}{n})$$

Pretendendo modelar o *heaping* , os autores propõem uma função de *heaping* independente de X e de T com a seguinte formulação:
Dado $q \in \mathbb{N}$ suficientemente grande, seja $\nu^{(l)}$ a probabilidade de uma duração ser prolongada l unidades e $\delta^{(l)}$ a probabilidade de uma duração ser reduzida l unidades, com $l = 1, 2, ..., q$ e $\xi = \Sigma_{l=1}^{q}(\nu^{(l)} + \delta^{(l)}) < 1$.

[6] Fraco consolo depois de tão duras provas!

Dadas as verdadeiras durações, $T_1, T_2, \ ..., T_n$, as durações observadas serão:

$$T_i^* = \begin{cases} T_i + q & \textit{com a probabilidade} & \nu^{(q)} \\ ... & " & ... \\ T_i + 1 & " & \nu^{(1)} \\ T_i & " & 1 - \xi \\ T_i - 1 & " & \delta^{(1)} \\ ... & " & ... \\ T_i - q & " & \delta^{(q)} \end{cases} \tag{2.3}$$

Mas neste caso como algum dos $\delta^{(l)}$ pode ser não nulo, algumas durações T_i^* poderão assumir valores negativos, pode por isso ser conveniente considerar um novo estimador:

$$\hat{\beta^{**}} = \frac{1}{\alpha} \ln(\frac{\sum_{i=1}^{n} (T_i^{**})^{\alpha}}{K})$$

onde $T_i^{**} = max\{0, T_i^*\}$ e K é o número de durações tais que $T_i^{**} > 0$

2.1.4.1 Resultados assimptóticos

Os estimadores $\hat{\beta^*}$ e $\hat{\beta^{**}}$ são em geral inconsistentes. Os autores demonstram que no caso precedente os seus enviesamentos assimptóticos são:

$\alpha = 1$ (caso exponencial)

$$\operatorname*{plim}_{n\to\infty} \hat{\beta^*} - \beta = ln\left(1 + \frac{\sum_{l=1}^{q} (\nu^{(l)} - \delta^{(l)}).l}{exp(\beta)}\right) \tag{2.4}$$

$$\underline{se} \quad \frac{\sum_{l=1}^{q} (\nu^{(l)} - \delta^{(l)}).l}{exp(\beta)} > -1$$

$\alpha = 2$ (distribuição *Rayleigh*)]

$$\operatorname*{plim}_{n\to\infty} \hat{\beta^*} - \beta = \frac{1}{2} ln\left(1 + \frac{\sum_{l=1}^{q} (\nu^{(l)} + \delta^{(l)}).l^2}{exp(2\beta)} + \frac{\sqrt{\pi} \sum_{l=1}^{q} (\nu^{(l)} - \delta^{(l)}).l}{exp(\beta)}\right) \tag{2.5}$$

$$\underline{se} \quad \frac{\sum_{l=1}^{q} (\nu^{(l)} + \delta^{(l)}).l^2}{exp(2\beta)} + \frac{\sqrt{\pi} \sum_{l=1}^{q} (\nu^{(l)} - \delta^{(l)}).l}{exp(\beta)} > -1$$

O enviesamento depende directamente da severidade do *heaping* e inversamente da magnitude de β.

Note-se que por construção $T_i^{**} \geq T_i^{*} \Rightarrow \hat{\beta^{**}} \geq \hat{\beta^{*}}$ logo

$$\operatorname*{plim}_{n\to\infty} \hat{\beta^{**}} \geq \operatorname*{plim}_{n\to\infty} \hat{\beta^{*}} \Rightarrow \operatorname*{plim}_{n\to\infty} \hat{\beta^{**}} - \beta \geq \operatorname*{plim}_{n\to\infty} \hat{\beta^{*}} - \beta$$

Sobressaem dois casos particulares:

1. Uma situação semelhante à encontrado no GOSEP por Kraus e Steiner (1998) em que as durações são sempre prolongadas, i. é o *heaping* é assimétrico positivo, $\delta^{(l)} = 0, \forall_{l\in\{1,2,\ldots,q\}}$ e nesse caso $\hat{\beta^{**}} = \hat{\beta^{*}}$

2. Uma situação semelhante à estudada por Torelli e Trivellato (1993), em que o *heaping* é praticamente simétrico, i. é, $\nu^{(l)} \simeq \delta^{(l)}$ $,\forall_{l\in\{1,2,\ldots,q\}}$. Nesse caso pelas equações (2.4) e (2.5) o estimador $\hat{\beta^{*}}$ é consistente no caso exponencial e inconsistente no caso *Rayleigh*, em linha com os resultados obtidos pelos autores italianos.

Conhecido o enviesamento de $\hat{\beta^{*}}$ é fácil encontrar um estimador consistente, ver (Augustin & Wolff, 2003, pg. 14).

2.1.4.2 Resultados para amostras finitas

Continuando a estudar dados de duração gerados por um PGD Weibull com tempo contínuo, os autores estudam por simulação de Monte Carlo o comportamento para amostras finitas ($N = 250$ e $N = 500$) dos estimadores *naïves* de máxima verosimilhança perante *heaping* (simétrico[7] e não simétrico[8]) fazendo variar o parâmetro α indicador da dependência temporal da duração ($\alpha = 1$, $\alpha = 1.4$ e $\alpha = 0.6$) e usando

[7] 11 % dos inícios das durações são *amontoadas*.

[8] Num primeiro caso 6 % dos inícios das durações e noutro 11 % dos inícios das durações são prolongados por *heaping* .

dois níveis para a média das durações ($\beta = 1$ e $\beta = 2$).

Em todos os casos os estimadores apresentam um enviesamento positivo, mais:

1. *Heaping* simétrico

   Em ambas as dimensões da amostra as diferenças entre os parâmetros estimados e o verdadeiro valor são estatisticamente nulas. O *heaping* simétrico não deteriora significativamente a qualidade dos estimadores

2. *Heaping* assimétrico positivo

   Neste caso os autores limitaram o estudo ao caso $\alpha = 1$ e a $N = 500$ tendo concluído que para $\beta = 2$ os estimadores estimam valores estatisticamente iguais ao verdadeiros valores dos parâmetros, mas para $\beta = 2$, embora por uma margem bastante escassa o estimador de β estima um valor $\hat{\beta}$ estatisticamente diferente do verdadeiro valor do parâmetro.

Finalmente, os autores estudam um possível efeito sazonal espúrio provocado por *heaping* , escolhem um PGD exponencial com $\beta = 1$, *heaping* assimétrico positivo (aproximadamente 6 % de observações prolongadas) e concluem ser significativamente diferentes de zero os coeficientes de algumas *dummy's* temporais incluídas como variáveis explicativas. Ou seja, até **baixos valores de *heaping* podem provocar um efeito sazonal espúrio**.

Capítulo 3

Estudo de simulação

3.1 Objectivos

É comum, nos modelos para dados de duração, recorrer a modelos de hazard proporcional em que os seus parâmetros são estimados pelo método de máxima verosimilhança (MV). Para uma descrição mais detalhada, Kiefer (1988).

Restringindo a nossa área de interesse aos modelos de hazard proporcional e em particular à distribuição Weibull, a partir dos trabalhos anteriores e perante os dois padrões de *heaping* mais observados[1], colocam-se-nos algumas questões:

1. A partir de que quantidade de *heaping* é de facto desaconselhável ignorar o *heaping* ?

2. Qual a performance dos estimadores de MV que ignoram o *heaping* na estimação dos parâmetros de um conjunto de variáveis explicativas maia alargado, em particular:

    - Variáveis dummy

    - Variáveis simétricas

[1] *Heaping* simétrico e *heaping* assimétrico positivo

- Variáveis assimétricas (assimetria positiva)

3. Qual a performance dos estimadores MV que ignoram o *heaping* perante a censura à direita?

Pretendendo estudar estas questões, faremos um estudo por simulação a descrever de seguida.

3.2 Processo gerador de dados(PGD)

O nosso objectivo será simular o efeito, na estimação dos parâmetros do modelo escolhido, do arredondamento de algumas durações geradas pelo PGD escolhido para durações mais "redondas".

As durações, T_i , serão geradas seguindo uma distribuição Weibull [2] de parâmetro α, tomando valores no conjunto $\{0.5,\ 1,\ 2\}$, reflectido os três tipos de dependência temporal característicos da Weibull e λ será a exponencial de uma combinação linear da *constante* e três variáveis explicativas, X_1 , X_2 e X_3, onde X_1 é uma variável dummy, X_2 é uma variável simétrica e X_3 assimétrica positiva, escolhendo-se a_1 a_2 e a_3 de forma a que o valor esperado das durações se situe entre 0.5 e 1 , i. é,

$$\lambda = exp(c + a_1X_1 + a_2X_2 + a_3X_3)$$

O *heaping* [3] terá o seguinte padrão:
As durações mais vulgares, entre 0 e 1, serão amontoadas para a décima mais próxima, durações menos esperadas, entre 1 e 3 serão amontoadas para o múltiplo de 0.5 mais próximo e as durações mais raras, maiores que 3, tomarão o valor mais próximo pertencente ao conjunto $\{\ 3,\ 4,\ ...,\ 9,\ 10,\ 15,\ 20,\ 30,\ 50,\ 100\}$. Ou seja, salvo indicação em contrário o nosso conjunto de *heaping* será:

$$H = \{\ 0.1,\ 0.2,\ ...,\ 0.9,\ 1,\ 1.5,\ 2,\ 2.5,\ 3,\ 4,\ ...,\ 9,\ 10,\ 15,\ 20,\ 30,\ 50,\ 100\}$$

[2] $f(t_i) = \alpha\lambda {t_i}^{\alpha-1} exp(-\lambda t_i^{\alpha})$

[3] Como $T > 0$, $0 \notin H$ o que introduz uma pequena assimetria positiva no padrão de *heaping* .

Na estimação será usado o método da máxima verosimilhança via a *package* Survival do software R [4].

É um facto bem conhecido que os estimadores de máxima verosimilhança são em geral enviesados mas consistentes. É portanto expectável algum afastamento entre os valores estimados e os verdadeiros valores dos parâmetros; tomaremos por isso, em geral, amostras suficientemente grandes, $N = 500$ e como critérios de qualidade o envisamento absoluto e um coeficiente de enviesamento relativo,

$$ER = \frac{\widehat{\mathbf{E}(\hat{\theta})} - \theta}{\theta}$$

que será apresentado em termos de percentagem e entre parênteses rectos.
Num primeiro estudo, com $N = 500$ e 5000 simulações estudaremos o desempenho dos estimadores de MV para percentagens de *heaping* crescentes e para diferentes combinações lineares das variáveis explicativas com especial enfoque no comportamento isolado de cada uma das variáveis.

De forma a tornar praticável e verosímil o padrão de *heaping* definido anteriormente, serão escolhidos valores para os parâmetros que afectam as variáveis explicativas de forma a que $E[T|\alpha,\ \lambda]$ se aproxime de 1 .

Num segundo estudo, com 5000 simulações procuraremos evidência empírica da esperada inconsistência dos estimadores para percentagens de *heaping* pequenas.

Finalmente, num terceiro estudo, ainda com $N = 500$ e 5000 simulações estudaremos o impacto da censura à direita na performance dos estimadores procurando comparar esses resultados com os obtidos para dados não censurados.

3.3 *Heaping* ignorado

Tal como descrito no capítulo anterior vamos analisar o impacto do *heaping* na estimação dos parâmetros do modelo escolhido, a saber:

[4] O código pode ser disponibilizado a quem mostrar interesse nele.

O dados são gerados seguindo uma distribuição Weibull com parâmetro de dependência temporal α e elasticidade λ, onde λ é a exponencial de uma combinação linear de variáveis aleatórias, $\lambda = exp(c + a_1X_1 + a_2X_2 + a_3X_3)$:
c é o termo autónomo do modelo;
X_1 é uma variável *dummy* tomando os valores 0 e 1 com média 0.5;
X_2 é uma variável *simétrica*, $X_2 \sim N(0, 1)$ e X_3 é uma variável *assimétrica*, $X_3 \sim \chi^2_{(2)}$.
Recorreu-se a 5000 simulações e a amostras de dimensão $N = 500$
A **título de exemplo** apresentamos a tabela 3.1 que reflecte a estimação do modelo Weibull com $\alpha = 0.5$ e $\lambda = exp(0.2+0.2X_1+0.2X_2+0.2X_3)$, entre parêntesis curvos o valor estimado do desvio padrão do estimador e entre parêntesis rectos o valor estimado do enviesamento relativo do estimador.

	Percentagem de *heaping*						
	0 %	5 %	10 %	25 %	50 %	75 %	100 %
$\hat{\alpha}$	0.503 (0.018) [0.7 %]	0.512 (0.018) [2.3 %]	0.520 (0.019) [4.1 %]	0.551 (0.020) [10.2 %]	0.602 (0.023) [22.3 %]	0.684 (0.026) [36.8 %]	0.769 (0.029) [53.7 %]
$\hat{c}$	0.201 (0.079) [0.6 %]	0.203 (0.080) [1.3 %]	0.202 (0.081) [1.0 %]	0.201 (0.084) [0.7 %]	0.187 (0.089) [-6.5 %]	0.155 (0.094) [-22. %]	0.105 (0.010) [-47.6 %]
$\hat{a}_1$	0.198 (0.091) [-0.9 %]	0.202 (0.090) [0.9 %]	0.202 (0.091) [0.8 %]	0.202 (0.094) [0.9 %]	0.209 (0.099) [4.7 %]	0.227 (0.011) [13.5 %]	0.247 (0.119) [23.5 %]
$\hat{a}_2$	0.202 (0.046) [1.1 %]	0.201 (0.046) [0.4 %]	0.200 (0.046) [0.1 %]	0.202 (0.047) [0.8 %]	0.209 (0.051) [4.5 %]	0.225 (0.055) [12.3 %]	0.245 (0.060) [22.5 %]
$\hat{a}_3$	0.203 (0.024) [1.3 %]	0.197 (0.025) [-1.3 %]	0.195 (0.025) [-2.7 %]	0.186 (0.026) [-7.1 %]	0.178 (0.025) [-10.9 %]	0.178 (0.024) [-10.8 %]	0.186 (0.024) [-6.8 %]

Tabela 3.1: Simulação para durações Weibull, $\alpha = 0.5$

Em termos globais, $\hat{\alpha}$ é tanto mais enviesado quanto maior for a percentagem de *heaping* . Em relação aos coeficientes da regressão, $\hat{c}$, $\hat{a_1}$, $\hat{a_2}$, $\hat{a_2}$ e $\hat{a_3}$, apresentam valores aceitáveis para percentagens de *heaping* até 25%, desde que os regressores associados tenham distribuições simétricas. Como veremos mais adiante a escolha de $\alpha = 0.5$ não é *inocente*.

Procurando clarificar os efeito, nos estimadores, de percentagens crescentes de *heaping* e por uma questão de simplicidade do estudo, simplificaremos o PGD do modelo, em particular, restringindo a uma só as variáveis explicativas com o intuito de discernir de forma mais rigorosa o comportamento dos estimadores.
O *heaping* variará entre 0% e 50%.

3.3.1 $\alpha = 0.5$

Segue-se os resultados da estimação para o parâmetro temporal $\alpha = 0.5$, entre parêntesis curvos uma estimativa do desvio padrão e entre parêntesis rectos uma estimativa do enviesamento relativo.
As tabelas que se seguem são um resumo das tabelas da secção A.1 do apêndice A:

Percentagem de *heaping*

	0%	5%	10%	30%	50%
X_1	$\hat{\alpha} = 0.502$	$\hat{\alpha} = 0.508$	$\hat{\alpha} = 0.514$	$\hat{\alpha} = 0.538$	$\hat{\alpha} = 0.564$
	(0.017)	(0.018)	(0.018)	(0.018)	(0.019)
	[0.4%]	[1.6%]	[2.8%]	[7.7%]	[12.9%]
	$\hat{a}_1 = 0.202$	$\hat{a}_1 = 0.204$	$\hat{a}_1 = 0.205$	$\hat{a}_1 = 0.208$	$\hat{a}_1 = 0.216$
	(0.092)	(0.092)	(0.091)	(0.096)	(0.098)
	[1%]	[1.8%]	[2.4%]	[3.9%]	[7.9%]
	$\alpha = 0.5$ – $a_1 = 0.2$				
	$\alpha = 0.5$ – $a_1 = 0.4$				
	$\hat{\alpha} = 0.502$	$\hat{\alpha} = 0.508$	$\hat{\alpha} = 0.515$	$\hat{\alpha} = 0.542$	$\hat{\alpha} = 0.572$
	(0.018)	(0.018)	(0.018)	(0.019)	(0.019)
	[0.5%]	[1.7%]	[3%]	[8.5%]	[14.4%]
	$\hat{a}_1 = 0.4$	$\hat{a}_1 = 0.405$	$\hat{a}_1 = 0.409$	$\hat{a}_1 = 0.419$	$\hat{a}_1 = 0.434$
	(0.092)	(0.092)	(0.092)	(0.097)	(0.099)
	[0.1%]	[1.4%]	[2.2%]	[4.7%]	[8.5%]

Tabela 3.2: Simulação para durações Weibull, $\alpha = 0.5$ - **parte superior** $\lambda = exp(0.2X_1)$ - **parte inferior** $\lambda = exp(0.4X_1)$.

O enviesamento relativo dos estimadores $\hat{a}_1$ e $\hat{\alpha}$ é positivo e cresce com a percentagem de *heaping* .

As estimativas de α e a_1 são aceitáveis até cerca de 20% de observações *amontoadas*.
Os casos $a_1 = 0.2$ e $a_1 = 0.4$ são muito semelhantes.

Percentagem de *heaping*

	0%	5%	10%	30%	50%
X_2	$\hat{\alpha} = 0.502$	$\hat{\alpha} = 0.508$	$\hat{\alpha} = 0.513$	$\hat{\alpha} = 0.535$	$\hat{\alpha} = 0.559$
	(0.018)	(0.018)	(0.018)	(0.018)	(0.019)
	[0.4%]	[1.5%]	[2.6%]	[7.1%]	[11.9%]
	$\hat{a}_2 = 0.201$	$\hat{a}_2 = 0.202$	$\hat{a}_2 = 0.203$	$\hat{a}_2 = 0.209$	$\hat{a}_2 = 0.215$
	(0.045)	(0.048)	(0.047)	(0.048)	(0.049)
	[0.7%]	[0.8%]	[1.6%]	[4.4%]	[7.4%]
	$\alpha = 0.5$ – $a_2 = 0.2$				
	$\alpha = 0.5$ – $a_2 = 0.4$				
	$\hat{\alpha} = 0.502$	$\hat{\alpha} = 0.508$	$\hat{\alpha} = 0.514$	$\hat{\alpha} = 0.537$	$\hat{\alpha} = 0.562$
	(0.018)	(0.018)	(0.018)	(0.019)	(0.019)
	[0.4%]	[1.5%]	[2.7%]	[7.4%]	[12.4%]
	$\hat{a}_2 = 0.402$	$\hat{a}_2 = 0.404$	$\hat{a}_2 = 0.406$	$\hat{a}_2 = 0.416$	$\hat{a}_2 = 0.428$
	(0.047)	(0.047)	(0.048)	(0.048)	(0.051)
	[0.6%]	[1%]	[1.4%]	[4%]	[7%]

Tabela 3.3: Simulação para durações Weibull, $\alpha = 0.5$ - **parte superior** $\lambda = exp(0.2X_2)$ - **parte inferior** $\lambda = exp(0.4X_2)$.

Padrão de estimação muito semelhante ao do caso anterior.

O enviesamento relativo dos estimadores $\hat{a}_2$ e $\hat{\alpha}$ é positivo e cresce com a percentagem de *heaping* .

As estimativas de α e a_2 são aceitáveis até cerca de 20% de observações *amontoadas*.

Os casos $a_2 = 0.2$ e $a_2 = 0.4$ são muito semelhantes. Na presença de *heaping* o enviesamento relativo de $\hat{\alpha}$ é sempre maior que o de $\hat{a}_2$.

Percentagem de *heaping*

	0%	5%	10%	30%	50%
X_3	$\hat{\alpha} = 0.502$	$\hat{\alpha} = 0.509$	$\hat{\alpha} = 0.517$	$\hat{\alpha} = 0.549$	$\hat{\alpha} = 0.585$
	(0.017)	(0.018)	(0.018)	(0.019)	(0.021)
	[0.3%]	[1.8%]	[3.3%]	[9.7%]	[17.1%]
	$\hat{a_3} = 0.201$	$\hat{a_3} = 0.2$	$\hat{a_3} = 0.197$	$\hat{a_3} = 0.193$	$\hat{a_3} = 0.191$
	(0.024)	(0.024)	(0.024)	(0.025)	(0.024)
	[0.7%]	[0%]	[-1.3%]	[-3.6%]	[-4.7%]
			$\alpha = 0.5$ – $a_3 = 0.2$		
			$\alpha = 0.5$ – $a_3 = 0.4$		
	$\hat{\alpha} = 0.502$	$\hat{\alpha} = 0.503$	$\hat{\alpha} = 0.506$	$\hat{\alpha} = 0.535$	$\hat{\alpha} = 0.580$
	(0.018)	(0.02)	(0.02)	(0.021)	(0.022)
	[0.4%]	[0.6%]	[1.1%]	[6.9%]	[15.9%]
	$\hat{a_3} = 0.402$	$\hat{a_3} = 0.372$	$\hat{a_3} = 0.346$	$\hat{a_3} = 0.285$	$\hat{a_3} = 0.256$
	(0.026)	(0.046)	(0.052)	(0.048)	(0.04)
	[0.5%]	[-7%]	[-13.4%]	[-28.8%]	[-36%]

Tabela 3.4: Simulação para durações Weibull, $\alpha = 0.5$ - **parte superior** $\lambda = exp(0.2X_3)$ - **parte inferior** $\lambda = exp(0.4X_3)$.

O enviesamento relativo do estimador $\hat{\alpha}$ é positivo e cresce com a percentagem de *heaping* . Os casos $a_3 = 0.2$ e $a_3 = 0.4$ são comparáveis mas não apresentam a mesma semelhança dos casos estudados anteriormente.

Na presença de *heaping* , o enviesamento relativo do estimador $\hat{a}_3$ é negativo e cresce, em valor absoluto, com a percentagem de *heaping* . As estimativas de α e a_3 não são aceitáveis praticamente para nenhuma das situações estudadas (sobretudo no caso $a_3 = 0.4$). No caso $a_3 = 0.4$ o enviesamento relativo é muito maior que no caso $a_3 = 0.2$ (mais uma vez em termos absolutos).

Como era de esperar o *heaping* enviesa a estimação dos parâmetros do modelo. A situação é semelhante no caso da variável dummy e da variável simétrica (tabelas 3.2

e 3.3 respectivamente), enviesamento relativo é positivo na estimação do parâmetro temporal e do coeficiente da variável explicativa. O caso da variável assimétrica (tabela 3.4) é marcadamente diferente dos dois anteriores, o enviesamento relativo é bastante maior e com padrão diferente do das duas outras variáveis explicativas. O parâmetro temporal continua a ser sobrestimado mas o coeficiente da variável explicativa é subestimado de forma particularmente severa no caso $a_3 = 0.4$.

3.3.2 $\alpha = 1$ (Exponencial)

Segue-se os resultados da estimação para o parâmetro temporal $\alpha = 1$, entre parêntesis curvos uma estimativa do desvio padrão e entre parêntesis rectos uma estimativa do enviesamento relativo.

As tabelas que se seguem são um resumo das tabelas da secção A.2 do apêndice A:

Percentagem de *heaping*

	0%	5%	10%	30%	50%
	$\hat{\alpha} = 1.004$	$\hat{\alpha} = 1.007$	$\hat{\alpha} = 1.01$	$\hat{\alpha} = 1.025$	$\hat{\alpha} = 1.038$
	(0.035)	(0.034)	(0.035)	(0.034)	(0.034)
	[0.4%]	[0.7%]	[1%]	[2.5%]	[3.8%]
	$\hat{a_1} = 0.201$	$\hat{a_1} = 0.2$	$\hat{a_1} = 0.202$	$\hat{a_1} = 0.204$	$\hat{a_1} = 0.207$
	(0.091)	(0.09)	(0.092)	(0.092)	(0.094)
	[0.7%]	[0.1%]	[1.2%]	[2.1%]	[3.7%]
			$\alpha = 1$ – $a_1 = 0.2$		
X_1					
			$\alpha = 1$ – $a_1 = 0.4$		
	$\hat{\alpha} = 1.004$	$\hat{\alpha} = 1.008$	$\hat{\alpha} = 1.012$	$\hat{\alpha} = 1.027$	$\hat{\alpha} = 1.043$
	(0.035)	(0.035)	(0.035)	(0.034)	(0.034)
	[0.4%]	[0.8%]	[1.2%]	[2.7%]	[4.3%]
	$\hat{a_1} = 0.4$	$\hat{a_1} = 0.404$	$\hat{a_1} = 0.404$	$\hat{a_1} = 0.408$	$\hat{a_1} = 0.415$
	(0.092)	(0.091)	(0.093)	(0.093)	(0.096)
	[0.1%]	[0.9%]	[1%]	[2%]	[3.8%]

Tabela 3.5: Simulação para durações Weibull (exponenciais), $\alpha = 1$ - **parte superior** $\lambda = exp(0.2X_1)$ - **parte inferior** $\lambda = exp(0.4X_1)$.

O enviesamento relativo dos estimadores, apesar de pouco significativo, aumenta com percentagens crescentes de *heaping* .

De notar porém que o *heaping* afecta pouco a estimação, inclusive quando 50% das observações são *amontoadas*, o enviesamento relativo mais pronunciado é o de $\hat{\alpha}$, apenas 4.3%.

Percentagem de *heaping*

	0%	5%	10%	30%	50%
X_2	$\hat{\alpha} = 1.003$	$\hat{\alpha} = 1.007$	$\hat{\alpha} = 1.01$	$\hat{\alpha} = 1.022$	$\hat{\alpha} = 1.035$
	(0.035)	(0.035)	(0.035)	(0.034)	(0.035)
	[0.3%]	[0.7%]	[1%]	[2.2%]	[3.5%]
	$\hat{a_2} = 0.202$	$\hat{a_2} = 0.202$	$\hat{a_2} = 0.204$	$\hat{a_2} = 0.204$	$\hat{a_2} = 0.206$
	(0.046)	(0.047)	(0.047)	(0.045)	(0.047)
	[0.9%]	[1.1%]	[1.8%]	[2.2%]	[2.8%]
	$\alpha = 1$ – $a_2 = 0.2$				
	$\alpha = 1$ – $a_2 = 0.4$				
	$\hat{\alpha} = 1.005$	$\hat{\alpha} = 1.007$	$\hat{\alpha} = 1.01$	$\hat{\alpha} = 1.024$	$\hat{\alpha} = 1.037$
	(0.035)	(0.035)	(0.036)	(0.035)	(0.034)
	[0.5%]	[0.7%]	[1%]	[2.4%]	[3.7%]
	$\hat{a_2} = 0.402$	$\hat{a_2} = 0.403$	$\hat{a_2} = 0.404$	$\hat{a_2} = 0.407$	$\hat{a_2} = 0.412$
	(0.047)	(0.047)	(0.048)	(0.048)	(0.049)
	[0.4%]	[0.6%]	[1%]	[1.9%]	[3.1%]

Tabela 3.6: Simulação para durações Weibull (exponenciais), $\alpha = 1$ - **parte superior** $\lambda = exp(0.2X_2)$ - **parte inferior** $\lambda = exp(0.4X_2)$.

O enviesamento relativo dos estimadores, ainda menos significativo que no caso da variável *dummy*, é positivo e aumenta com percentagens crescentes de *heaping* .
Mais uma vez o *heaping* afecta pouco a estimação, quando 50% das observações são *amontoadas*, o enviesamento relativo mais pronunciado é o de $\hat{\alpha}$, 3.73%.
Na ausência de dependência temporal negativa o *heaping* não parece ser um factor a ter em conta se os regressores seguirem distribuições simétricas. Vejamos se o caso $\alpha = 2$ confirma essa tendência.

Percentagem de *heaping*

	0%	5%	10%	30%	50%
X_3	$\hat{\alpha} = 1.004$	$\hat{\alpha} = 1.009$	$\hat{\alpha} = 1.014$	$\hat{\alpha} = 1.034$	$\hat{\alpha} = 1.055$
	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)
	[0.4%]	[0.9%]	[1.4%]	[3.4%]	[5.5%]
	$\hat{a_3} = 0.202$	$\hat{a_3} = 0.202$	$\hat{a_3} = 0.202$	$\hat{a_3} = 0.202$	$\hat{a_3} = 0.204$
	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)
	[0.8%]	[1%]	[0.9%]	[1.2%]	[2%]
			$\alpha = 1$ – $a_3 = 0.2$		
			$\alpha = 1$ – $a_3 = 0.4$		
	$\hat{\alpha} = 1.005$	$\hat{\alpha} = 1.009$	$\hat{\alpha} = 1.014$	$\hat{\alpha} = 1.036$	$\hat{\alpha} = 1.065$
	(0.035)	(0.037)	(0.036)	(0.038)	(0.037)
	[0.5%]	[0.9%]	[1.4%]	[3.6%]	[6.5%]
	$\hat{a_3} = 0.402$	$\hat{a_3} = 0.392$	$\hat{a_3} = 0.386$	$\hat{a_3} = 0.356$	$\hat{a_3} = 0.338$
	(0.027)	(0.035)	(0.038)	(0.044)	(0.044)
	[0.6%]	[−1.9%]	[−3.6%]	[−10.9%]	[−15.4%]

Tabela 3.7: Simulação para durações Weibull (exponenciais), $\alpha = 1$ - **parte superior** $\lambda = exp(0.2X_3)$ - **parte inferior** $\lambda = exp(0.4X_3)$.

O enviesamento relativo de $\hat{\alpha}$, é positivo e aumenta com percentagens crescentes de *heaping* . As duas situações, $a_3 = 0.2$ e $a_3 = 0.4$, são semelhantes.

O enviesamento relativo de $\hat{a}_3$ é positivo e pouco significativo no caso $a_3 = 0.2$ mas no caso $a_3 = 0.4$ é negativo e é aceitável apenas para percentagens de *heaping* inferiores a 20%.

No caso exponencial o padrão do enviesamento relativo é o mesmo do caso $\alpha = 0.5$ embora bastante menos pronunciado, i. é, o efeito do *heaping* enviesa da mesma forma a estimação mas conduz a enviesamentos menos pronunciados. Note-se, em particular, e mais uma vez o caso da variável assimétrica (tabela 3.7) onde o enviesamento relativo é maior. No caso exponencial se os regressores seguirem distribuições simétricas o *heaping* não afecta significativamente a estimação, o que está em linha com os resultados deduzidos por Augustin e Wolff (2003) para $\lambda = c$. Verifiquemos o caso seguinte, $\alpha = 2$, ou seja, dependência temporal positiva.

3.3.3 $\alpha = 2$

Segue-se os resultados da estimação para o parâmetro temporal $\alpha = 2$, entre parêntesis curvos uma estimativa do desvio padrão e entre parêntesis rectos uma estimativa do enviesamento relativo.

As tabelas que se seguem são um resumo das tabelas da secção A.3 do apêndice A:

Percentagem de *heaping*

	0%	5%	10%	30%	50%
X_1	$\hat{\alpha} = 2.007$	$\hat{\alpha} = 2.006$	$\hat{\alpha} = 2.004$	$\hat{\alpha} = 1.999$	$\hat{\alpha} = 1.993$
	(0.07)	(0.07)	(0.069)	(0.07)	(0.068)
	[0.3%]	[0.3%]	[0.2%]	[−0.1%]	[−0.3%]
	$\hat{a_1} = 0.2$	$\hat{a_1} = 0.2$	$\hat{a_1} = 0.2$	$\hat{a_1} = 0.203$	$\hat{a_1} = 0.2$
	(0.091)	(0.091)	(0.09)	(0.091)	(0.091)
	[0.1%]	[0.2%]	[0.2%]	[1.4%]	[0.1%]
			$\alpha = 2$ - $a_1 = 0.2$		
			$\alpha = 2$ - $a_1 = 0.4$		
	$\hat{\alpha} = 2.009$	$\hat{\alpha} = 2.005$	$\hat{\alpha} = 2.006$	$\hat{\alpha} = 2.002$	$\hat{\alpha} = 1.996$
	(0.07)	(0.07)	(0.07)	(0.069)	(0.069)
	[0.5%]	[0.3%]	[0.3%]	[0.1%]	[−0.2%]
	$\hat{a_1} = 0.403$	$\hat{a_1} = 0.399$	$\hat{a_1} = 0.404$	$\hat{a_1} = 0.401$	$\hat{a_1} = 0.404$
	(0.092)	(0.092)	(0.092)	(0.092)	(0.093)
	[0.8%]	[−0.3%]	[1%]	[0.3%]	[1%]

Tabela 3.8: Simulação para durações Weibull, $\alpha = 2$ - **parte superior** $\lambda = exp(0.2X_1)$ - **parte inferior** $\lambda = exp(0.4X_1)$.

O padrão de *heaping* em estudo virtualmente não aumenta o enviesamento relativo dos estimadores quando a variável explicativa é uma variável dummy e $\alpha = 2$. Verifiquemos se o mesmo ocorre com as outras variáveis.

Percentagem de *heaping*

	0%	5%	10%	30%	50%
	$\hat{\alpha} = 2.008$	$\hat{\alpha} = 2.008$	$\hat{\alpha} = 2.005$	$\hat{\alpha} = 1.996$	$\hat{\alpha} = 1.991$
	(0.07)	(0.071)	(0.071)	(0.068)	(0.068)
	[0.4%]	[0.4%]	[0.2%]	[-0.2%]	[-0.4%]
	$\hat{a_2} = 0.201$	$\hat{a_2} = 0.201$	$\hat{a_2} = 0.2$	$\hat{a_2} = 0.2$	$\hat{a_2} = 0.201$
	(0.045)	(0.045)	(0.047)	(0.045)	(0.046)
	[0.7%]	[0.7%]	[0.2%]	[0.2%]	[0.6%]
X_2			$\alpha = 2$ – $a_2 = 0.2$		
			$\alpha = 2$ – $a_2 = 0.4$		
	$\hat{\alpha} = 2.006$	$\hat{\alpha} = 2.006$	$\hat{\alpha} = 2.006$	$\hat{\alpha} = 1.999$	$\hat{\alpha} = 1.993$
	(0.07)	(0.07)	(0.071)	(0.069)	(0.068)
	[0.3%]	[0.3%]	[0.3%]	[0%]	[-0.4%]
	$\hat{a_2} = 0.401$	$\hat{a_2} = 0.4$	$\hat{a_2} = 0.401$	$\hat{a_2} = 0.401$	$\hat{a_2} = 0.399$
	(0.047)	(0.048)	(0.048)	(0.048)	(0.048)
	[0.4%]	[0%]	[0.4%]	[0.4%]	[-0.2%]

Tabela 3.9: Simulação para durações Weibull, $\alpha = 2$ - **parte superior** $\lambda = exp(0.2X_2)$ - **parte inferior** $\lambda = exp(0.4X_2)$.

Os resultados são muito semelhantes aos obtidos no caso anterior. Dir-se-ia que *algum heaping* até melhora o desempenho dos estimadores de MV no que toca ao enviesamento relativo. Verifiquemos o, até agora "desviante", caso da variável assimétrica.

Percentagem de *heaping*

	0%	5%	10%	30%	50%
	$\hat{\alpha} = 2.008$	$\hat{\alpha} = 2.008$	$\hat{\alpha} = 2.006$	$\hat{\alpha} = 2.005$	$\hat{\alpha} = 2.003$
	(0.071)	(0.07)	(0.071)	(0.07)	(0.071)
	[0.4%]	[0.4%]	[0.3%]	[0.2%]	[0.1%]
	$\hat{a_3} = 0.202$	$\hat{a_3} = 0.202$	$\hat{a_3} = 0.202$	$\hat{a_3} = 0.202$	$\hat{a_3} = 0.201$
	(0.024)	(0.025)	(0.024)	(0.024)	(0.024)
	[0.8%]	[1.1%]	[0.8%]	[0.8%]	[0.5%]
X_3			$\alpha = 2$ - $a_3 = 0.2$		
			$\alpha = 2$ - $a_3 = 0.4$		
	$\hat{\alpha} = 2.009$	$\hat{\alpha} = 2.008$	$\hat{\alpha} = 2.008$	$\hat{\alpha} = 2.011$	$\hat{\alpha} = 2.012$
	(0.071)	(0.069)	(0.069)	(0.069)	(0.071)
	[0.4%]	[0.4%]	[0.4%]	[0.5%]	[0.6%]
	$\hat{a_3} = 0.403$	$\hat{a_3} = 0.401$	$\hat{a_3} = 0.401$	$\hat{a_3} = 0.396$	$\hat{a_3} = 0.393$
	(0.027)	(0.027)	(0.028)	(0.029)	(0.03)
	[0.7%]	[0.4%]	[0.1%]	[−1%]	[−1.8%]

Tabela 3.10: Simulação para durações Weibull, $\alpha = 2$ - **parte superior** $\lambda = exp(0.2X_3)$ - **parte inferior** $\lambda = exp(0.4X_3)$.

Este caso é semelhante aos dois anteriores. O padrão de *heaping* em estudo virtualmente não aumenta o enviesamento relativo dos estimadores de MV para a dependência temporal considerada.
Apenas no caso $a_3 = 0.4$ e para grandes percentagens de *heaping* se nota algum aumento do enviesamento relativo mas nada comparável ao obtidos para as dependências temporais estudadas anteriormente.

No caso $\alpha = 2$ o padrão de *heaping* escolhido neste estudo praticamente não interfere com a estimação, pode até dizer-se que nalguns casos "melhora"a qualidade dos estimadores (tabelas 3.8, 3.9 e 3.10)!

Em conclusão, ignorar o *heaping* (com o padrão escolhido neste estudo) não conduz forçosamente a uma degradação da performance dos estimadores. A performance está sobretudo dependente da dimensão de α, quanto maior for, menor será o impacto do *heaping* .

Quanto à questão formulada, "A partir de que quantidade de *heaping* é de facto desaconselhável ignorar o *heaping* ? ", a resposta deverá ser:

"Depende essencialmente do parâmetro da dependência temporal ".

Uma última nota para a variância estimada dos estimadores em estudo; estas apresentam dimensões semelhantes para diferentes percentagens de *heaping* o que reforça a comparabilidade das estimativas [5].

3.3.4 Variando α

Naturalmente os resultados anteriores sugerem a continuação do estudo.

Procurando reforçar a suspeita de que a dimensão de α tem um papel determinante na magnitude do enviesamento relativo dos estimadores de MV, nesta subsecção faremos variar α de 0.3 até 2.0.

Como nos resultados obtidos até aqui a perturbação mais severa ocorreu na variável assimétrica no caso $\lambda = exp(0.4X_3)$, será este o valor escolhido para λ.

De notar que $\alpha = 0.1$ e $\alpha = 0.2$, foram também estudados mas, conduzem a valores demasiado elevados do valor esperado das durações, $E[T|\alpha, \lambda] \gg 1$ e atendendo ao conjunto de *heaping* escolhido, $H = \{..., 10, 15, 20, 30, 50, 100\}$, inviabilizam muitas vezes, por singularidade, a estimação ou descaracterizam os resultados uma vez que tornam o padrão de *heaping* demasiado inverosímil [6].

A tabela seguinte, tabela 3.11, ilustra a evolução com a dimensão de α do enviesamento relativo na estimação.

[5] O erro quadrático médio de um estimador $\hat{\beta}$ é $EQM(\hat{\beta}) = Var(\hat{\beta}) + [Env(\hat{\beta})]^2$.

[6] Em muitos casos $t_i \gg t_i^*$.

α					
0.3	0.4	0.5	0.6	0.7	0.8
$\hat{\alpha} = 0.366$ [21.9%] $\hat{a}_3 = 0.210$ [−47.5%]	$\hat{\alpha} = 0.475$ [18.8%] $\hat{a}_3 = 0.234$ [−41.5%]	$\hat{\alpha} = 0.580$ [16.0%] $\hat{a}_3 = 0.256$ [−36.0%]	$\hat{\alpha} = 0.681$ [13.5%] $\hat{a}_3 = 0.277$ [−30.8%]	$\hat{\alpha} = 0.779$ [11.3%] $\hat{a}_3 = 0.294$ [−26.5%]	$\hat{\alpha} = 0.875$ [9.4%] $\hat{a}_3 = 0.311$ [−22.2%]
α					
0.9	1.0	1.1	1.2	1.3	1.4
$\hat{\alpha} = 0.970$ [7.8%] $\hat{a}_3 = 0.325$ [−18.8%]	$\hat{\alpha} = 1.065$ [6.5%] $\hat{a}_3 = 0.338$ [−15.6%]	$\hat{\alpha} = 1.161$ [5.5%] $\hat{a}_3 = 0.350$ [−12.5%]	$\hat{\alpha} = 1.254$ [4.5%] $\hat{a}_3 = 0.360$ [−10.1%]	$\hat{\alpha} = 1.347$ [3.6%] $\hat{a}_3 = 0.367$ [−8.3%]	$\hat{\alpha} = 1.442$ [3.0%] $\hat{a}_3 = 0.375$ [−6.3%]
α					
1.5	1.6	1.7	1.8	1.9	2.0
$\hat{\alpha} = 1.535$ [2.3%] $\hat{a}_3 = 0.378$ [−5.6%]	$\hat{\alpha} = 1.630$ [1.9%] $\hat{a}_3 = 0.382$ [−4.4%]	$\hat{\alpha} = 1.725$ [1.4%] $\hat{a}_3 = 0.386$ [−3.5%]	$\hat{\alpha} = 1.821$ [1.2%] $\hat{a}_3 = 0.390$ [−2.5%]	$\hat{\alpha} = 1.915$ [0.8%] $\hat{a}_3 = 0.391$ [−2.3%]	$\hat{\alpha} = 2.012$ [0.6%] $\hat{a}_3 = 0.393$ [−1.8%]

Tabela 3.11: Simulação para durações Weibull, α entre 0.3 e 2.0 , $\lambda = exp(0.4X_3)$ e 50% de *heaping* .

O enviesamento relativo positivo de $\hat{\alpha}$ decresce com a dimensão de α tornando-se aceitável para dependências temporais significativamente positivas [7]. O enviesamento relativo negativo na estimação de a_3 decresce, em valor absoluto, com a dimensão de α tornando-se aceitável para dependências temporais significativamente próximas de 2.

[7] Maiores que 1

3.4 *Heaping* e inconsistência dos estimadores de máxima verosimilhança

Nesta secção avaliamos o comportamento em amostras finitas dos estimadores estudados até aqui.

Da literatura precedente (Augustin & Wolff, 2003) o *heaping* conduz à inconsistência dos estimadores de máxima verosimilhança quando o modelo Weibull tem a elasticidade definida por $\lambda = exp(c)$.

Mais uma vez, alargando o conjunto de variáveis explicativas como até aqui, estudaremos o enviesamento relativo dos estimadores de MV fazendo variar a dimensão da amostra, N tomará os valores 100, 400, 1600 e 6400. Seguem-se os valores obtidos da simulação para o modelo Weibull, com $\alpha = 0.5$ e 5000 réplicas, tabelas 3.12 e 3.13.

A primeira tabela refere-se a $\lambda = exp(0.2X_1)$, a segunda a $\lambda = exp(0.2X_2)$ e $\lambda = exp(0.2X_3)$.

		Percentagem de *heaping*			
	N	0%	5%	10%	50%
$\lambda = exp(0.2X_1)$	100	$\hat{\alpha} = 0.510$	$\hat{\alpha} = 0.516$	$\hat{\alpha} = 0.522$	$\hat{\alpha} = 0.575$
		$\hat{a}_1 = 0.203$	$\hat{a}_1 = 0.206$	$\hat{a}_1 = 0.206$	$\hat{a}_1 = 0.222$
	400	$\hat{\alpha} = 0.503$	$\hat{\alpha} = 0.508$	$\hat{\alpha} = 0.514$	$\hat{\alpha} = 0.565$
		$\hat{a}_1 = 0.199$	$\hat{a}_1 = 0.205$	$\hat{a}_1 = 0.204$	$\hat{a}_1 = 0.216$
	1600	$\hat{\alpha} = 0.500$	$\hat{\alpha} = 0.506$	$\hat{\alpha} = 0.512$	$\hat{\alpha} = 0.563$
		$\hat{a}_1 = 0.200$	$\hat{a}_1 = 0.201$	$\hat{a}_1 = 0.203$	$\hat{a}_1 = 0.215$
	6400	$\hat{\alpha} = 0.500$	$\hat{\alpha} = 0.506$	$\hat{\alpha} = 0.512$	$\hat{\alpha} = 0.562$
		$\hat{a}_1 = 0.200$	$\hat{a}_1 = 0.202$	$\hat{a}_1 = 0.202$	$\hat{a}_1 = 0.215$

Tabela 3.12: Simulação para durações Weibull, $\alpha = 0.5$

	Percentagem de *heaping*				
	N	0%	5%	10%	50%
$\lambda = exp(0.2X_2)$	100	$\hat{\alpha} = 0.510$	$\hat{\alpha} = 0.516$	$\hat{\alpha} = 0.521$	$\hat{\alpha} = 0.568$
		$\hat{a}_2 = 0.207$	$\hat{a}_2 = 0.206$	$\hat{a}_2 = 0.207$	$\hat{a}_2 = 0.221$
	400	$\hat{\alpha} = 0.503$	$\hat{\alpha} = 0.508$	$\hat{\alpha} = 0.513$	$\hat{\alpha} = 0.560$
		$\hat{a}_2 = 0.200$	$\hat{a}_2 = 0.202$	$\hat{a}_2 = 0.203$	$\hat{a}_2 = 0.214$
	1600	$\hat{\alpha} = 0.500$	$\hat{\alpha} = 0.506$	$\hat{\alpha} = 0.511$	$\hat{\alpha} = 0.558$
		$\hat{a}_2 = 0.200$	$\hat{a}_2 = 0.202$	$\hat{a}_2 = 0.203$	$\hat{a}_2 = 0.214$
	6400	$\hat{\alpha} = 0.500$	$\hat{\alpha} = 0.505$	$\hat{\alpha} = 0.511$	$\hat{\alpha} = 0.557$
		$\hat{a}_2 = 0.200$	$\hat{a}_2 = 0.201$	$\hat{a}_2 = 0.202$	$\hat{a}_2 = 0.214$
$\lambda = exp(0.2X_3)$	100	$\hat{\alpha} = 0.509$	$\hat{\alpha} = 0.518$	$\hat{\alpha} = 0.525$	$\hat{\alpha} = 0.597$
		$\hat{a}_3 = 0.209$	$\hat{a}_3 = 0.206$	$\hat{a}_3 = 0.204$	$\hat{a}_3 = 0.201$
	400	$\hat{\alpha} = 0.502$	$\hat{\alpha} = 0.510$	$\hat{\alpha} = 0.517$	$\hat{\alpha} = 0.586$
		$\hat{a}_3 = 0.202$	$\hat{a}_3 = 0.201$	$\hat{a}_3 = 0.198$	$\hat{a}_3 = 0.191$
	1600	$\hat{\alpha} = 0.501$	$\hat{\alpha} = 0.508$	$\hat{\alpha} = 0.515$	$\hat{\alpha} = 0.583$
		$\hat{a}_3 = 0.201$	$\hat{a}_3 = 0.198$	$\hat{a}_3 = 0.196$	$\hat{a}_3 = 0.188$
	6400	$\hat{\alpha} = 0.500$	$\hat{\alpha} = 0.507$	$\hat{\alpha} = 0.515$	$\hat{\alpha} = 0.582$
		$\hat{a}_3 = 0.200$	$\hat{a}_3 = 0.198$	$\hat{a}_3 = 0.195$	$\hat{a}_3 = 0.186$

Tabela 3.13: Simulação para durações Weibull, $\alpha = 0.5$

Se o *heaping* não levasse à inconsistência dos estimadores estes convergiriam à taxa $\sqrt{N}$. Note-se que a convergência é evidente para o caso de o *heaping* ser de 0%, mas isso, é claro, corresponde a **ausência de *heaping*** .

Quando a dimensão da amostra passa de $N = 100$ para $N = 400$ o enviesamento relativo dos estimadores de α diminui sensivelmente em todos os casos mas a melhoria já não é tão evidente de $N = 400$ para $N = 1600$ e de $N = 1600$ para $N = 6400$ na prática não há nenhuma melhoria.

O comportamento dos estimadores dos parâmetros das variáveis explicativas é ainda mais notável, mais uma vez há dois casos a considerar, variáveis explicativas com distribuição simétrica e variáveis explicativas com distribuição assimétrica. No caso simétrico, quando a dimensão da amostra passa de $N = 100$ para $N = 400$ o enviesamento relativo diminui sensivelmente em todos os casos mas, na presença de *heaping* , a melhoria já não é tão evidente de $N = 400$ para $N = 1600$ e no caso da variável X_2, normal, não há melhoria; de $N = 1600$ para $N = 6400$ na prática não há nenhuma melhoria, sobretudo para valores de *heaping* significativos (ver 50%).

No caso assimétrico, com apenas 10% de *heaping* , amostras maiores levam a maior enviesamento relativo do estimador $\hat{a}_3$ o que é um fortíssimo sinal de inconsistência do estimador em estudo.

No essencial estes resultados reforçam os obtidos por Augustin e Wolff (2003) fornecendo evidência empírica da inconsistência dos estimadores de MV na presença de *heaping* simétrico para modelos Weibull com dependência temporal negativa.

3.5 Dados censurados

Nesta secção estudaremos o impacto da censura no modelo estudado nas secções precedentes.

Em dados de duração a censura aparece associada ao facto dos eventos estudados nem sempre se completarem durante o decorrer do estudo. Neste tipo de censura as durações maiores têm uma maior probabilidade de ser censuradas. Uma forma de lidar com a censura na estimação (por MV) de modelos paramétricos é assumir que durações completas e durações censuradas contribuem de forma diferenciada para a verosimilhança do modelo [8] . No software usado, R, tal facto é facilmente acomodado atribuindo o $status = 1$, para durações completas e o $status = 0$, para durações censuradas.

As durações, T_i, serão geradas, tal com até aqui, seguindo uma distribuição Weibull de parâmetros $\alpha \in \{0.5, 1, 2\}$ e $\lambda \in \{e^{0.4X_1}, e^{0.4X_2}, e^{0.4X_3}\}$; onde X_1 é uma variável

[8] No primeiro caso conhecemos $f(t_i)$ e no segundo $S(t_i)$

dummy com 50% de $1's$, $X_2 \sim N(0,1)$ e $X_3 \sim \chi^2_{(2)}$.
Na modelação da censura seguiremos a proposta de Horowitz e Neuman (1989); a variável indicadora de censura, W, seguirá uma distribuição uniforme de mínimo $m = 10^{-6}$ e máximo M, com M escolhido caso a caso para gerar a percentagem de censura pretendida em cada situação estudada [9]. As durações "observadas"serão então $T^* = min\{T, W\}$.

Pretendendo isolar os efeitos da censura em cada um dos tipos de variável explicativa as subsecções seguintes serão indexadas a X_1, X_2 e X_3 respectivamente. A censura será num primeiro caso fraca, 10% de observações censuradas, num segundo caso moderada, 30% de observações censuradas e finalmente num terceiro caso severa, 50% de observações censuradas.
Os resultados obtidos estão resumidos e comentados nas subsecções que se seguem, entre parêntesis curvos uma estimativa do desvio padrão dos estimadores e entre parêntesis rectos uma estimativa do enviesamento relativo dos estimadores .

3.5.1 Dummy como variável explicativa

Nesta subsecção, $\lambda = exp(0.4X_1)$, os resultados das três tabelas que se seguem **serão comparados** com os da parte inferior das tabelas das páginas 28, 32 e 35 [10].

[9] Como as durações censuradas têm de ser estritamente positivas m não pode tomar o valor 0, a escolha de 10^{-6} serve esse propósito.

[10] Variável X_1, *heaping* ignorado com $\alpha = 0.5$, $\alpha = 1$ e $\alpha = 2$, respectivamente.

		Percentagem de heaping			
		0 %	10 %	20 %	50 %
$\lambda = exp(0.4X_1)$	$\alpha = 0.5$	$\hat{\alpha} = 0.502$	$\hat{\alpha} = 0.517$	$\hat{\alpha} = 0.533$	$\hat{\alpha} = 0.585$
		(0.02) [0.5 %]	(0.02) [3.4 %]	(0.02) [6.5 %]	(0.02) [17.1 %]
		$\hat{a}_1 = 0.399$	$\hat{a}_1 = 0.408$	$\hat{a}_1 = 0.409$	$\hat{a}_1 = 0.428$
		(0.10) [-0.3 %]	(0.10) [2.0 %]	(0.10) [2.3 %]	(0.11) [6.9 %]
	$\alpha = 1$	$\hat{\alpha} = 1.004$	$\hat{\alpha} = 1.012$	$\hat{\alpha} = 1.022$	$\hat{\alpha} = 1.048$
		(0.04) [0.4 %]	(0.04) [1.2 %]	(0.04) [2.2 %]	(0.04) [4.8 %]
		$\hat{a}_1 = 0.401$	$\hat{a}_1 = 0.404$	$\hat{a}_1 = 0.405$	$\hat{a}_1 = 0.415$
		(0.09) [0.2 %]	(0.09) [0.9 %]	(0.10) [1.3 %]	(0.10) [3.8 %]
	$\alpha = 2$	$\hat{\alpha} = 2.008$	$\hat{\alpha} = 2.006$	$\hat{\alpha} = 2.005$	$\hat{\alpha} = 2.000$
		(0.07) [0.4 %]	(0.07) [0.3 %]	(0.07) [0.2 %]	(0.07) [0.0 %]
		$\hat{a}_1 = 0.402$	$\hat{a}_1 = 0.399$	$\hat{a}_1 = 0.403$	$\hat{a}_1 = 0.402$
		(0.10) [0.5 %]	(0.10) [-0.3 %]	(0.10) [0.6 %]	(0.10) [0.4 %]

Tabela 3.14: Simulação para durações Weibull, $\lambda = exp(0.4X_1)$ e 10 % de observações censuradas

A influência de censura fraca não tem uma interpretação óbvia sobre a estimação quer de α quer de a_1 mas tendemos a aceitar que as pequenas diferenças de desempenho se devam a causas aleatórias, i. é, a presença de censura fraca não afecta a performance dos estimadores de MV quando comparados com os estimadores equivalentes na ausência de censura. Contudo os resultados que se seguem não reforçam esta interpretação.

		Percentagem de heaping			
		0 %	10 %	20 %	50 %
$\lambda = exp(0.4X_1)$	$\alpha = 0.5$	$\hat{\alpha} = 0.502$	$\hat{\alpha} = 0.522$	$\hat{\alpha} = 0.544$	$\hat{\alpha} = 0.623$
		(0.02) [0.3 %]	(0.02) [4.5 %]	(0.03) [8.7 %]	(0.03) [24.5 %]
		$\hat{a}_1 = 0.402$	$\hat{a}_1 = 0.402$	$\hat{a}_1 = 0.405$	$\hat{a}_1 = 0.413$
		(0.11) [0.4 %]	(0.11) [0.5 %]	(0.11) [1.3 %]	(0.12) [3.2 %]
	$\alpha = 1$	$\hat{\alpha} = 1.005$	$\hat{\alpha} = 1.015$	$\hat{\alpha} = 1.028$	$\hat{\alpha} = 1.065$
		(0.04) [0.5 %]	(0.04) [1.5 %]	(0.04) [2.8 %]	(0.04) [6.5 %]
		$\hat{a}_1 = 0.404$	$\hat{a}_1 = 0.406$	$\hat{a}_1 = 0.405$	$\hat{a}_1 = 0.413$
		(0.11) [0.9 %]	(0.11) [1.4 %]	(0.11) [1.3 %]	(0.11) [3.4 %]
	$\alpha = 2$	$\hat{\alpha} = 2.008$	$\hat{\alpha} = 2.010$	$\hat{\alpha} = 2.010$	$\hat{\alpha} = 2.008$
		(0.08) [0.4 %]	(0.08) [0.5 %]	(0.08) [0.5 %]	(0.08) [0.4 %]
		$\hat{a}_1 = 0.402$	$\hat{a}_1 = 0.405$	$\hat{a}_1 = 0.403$	$\hat{a}_1 = 0.401$
		(0.11) [0.4 %]	(0.11) [1.3 %]	(0.11) [0.9 %]	(0.11) [0.4 %]

Tabela 3.15: Simulação para durações Weibull, $\lambda = exp(0.4X_1)$ e 30 % de observações censuradas

Esta tabela começa a revelar um padrão curioso, por um lado o enviesamento relativo do estimador de α aumenta (relativamente à situação de ausência de censura) mas o enviesamento relativo do estimador de a_1 diminui.

		Percentagem de heaping			
		0 %	10 %	20 %	50 %
$\lambda = exp(0.4X_1)$	$\alpha = 0.5$	$\hat{\alpha} = 0.502$ (0.03) [0.3 %]	$\hat{\alpha} = 0.530$ (0.03) [6.1 %]	$\hat{\alpha} = 0.564$ (0.03) [12.7 %]	$\hat{\alpha} = 0.693$ (0.04) [38.7 %]
		$\hat{a}_1 = 0.401$ (0.13) [0.3 %]	$\hat{a}_1 = 0.396$ (0.13) [-1.1 %]	$\hat{a}_1 = 0.404$ (0.13) [0.9 %]	$\hat{a}_1 = 0.403$ (0.13) [0.9 %]
	$\alpha = 1$	$\hat{\alpha} = 1.005$ (0.05) [0.5 %]	$\hat{\alpha} = 1.024$ (0.05) [2.4 %]	$\hat{\alpha} = 1.042$ (0.05) [4.2 %]	$\hat{\alpha} = 1.103$ (0.05) [10.3 %]
		$\hat{a}_1 = 0.403$ (0.13) [0.7 %]	$\hat{a}_1 = 0.404$ (0.13) [1.0 %]	$\hat{a}_1 = 0.404$ (0.13) [1.1 %]	$\hat{a}_1 = 0.410$ (0.13) [2.4 %]
	$\alpha = 2$	$\hat{\alpha} = 2.008$ (0.10) [0.4 %]	$\hat{\alpha} = 2.010$ (0.10) [0.5 %]	$\hat{\alpha} = 2.015$ (0.10) [0.7 %]	$\hat{\alpha} = 2.019$ (0.10) [0.9 %]
		$\hat{a}_1 = 0.401$ (0.13) [0.2 %]	$\hat{a}_1 = 0.404$ (0.13) [0.9 %]	$\hat{a}_1 = 0.402$ (0.13) [0.6 %]	$\hat{a}_1 = 0.404$ (0.13) [0.9 %]

Tabela 3.16: Simulação para durações Weibull, $\lambda = exp(0.4X_1)$ e 50 % de observações censuradas

Esta tabela reforça de forma evidente o padrão observado na anterior, repare-se que o enviesamento relativo do estimador de α aumenta mas, mesmo no caso de 50% de observações *amontoadas*, o enviesamento relativo de $\hat{a}_1$ é quase nulo.

Verifiquemos o que ocorre no caso da variável simétrica.

3.5.2 Variável simétrica como variável explicativa

Nesta subsecção, $\lambda = exp(0.4X_2)$. Os resultados das três tabelas que se seguem serão comparados com os da parte inferior das tabelas das páginas 29, 33 e 36 [11].

[11] Variável X_2, *heaping* ignorado com $\alpha = 0.5$, $\alpha = 1$ e $\alpha = 2$, respectivamente.

		Percentagem de heaping			
		0 %	10 %	20 %	50 %
$\lambda = exp(0.4X_2)$	$\alpha = 0.5$	$\hat{\alpha} = 0.502$ (0.02) [0.4 %] $\hat{a}_2 = 0.403$ (0.5) [0.7 %]	$\hat{\alpha} = 0.518$ (0.02) [3.5 %] $\hat{a}_2 = 0.406$ (0.05) [1.5 %]	$\hat{\alpha} = 0.534$ (0.02) [6.7 %] $\hat{a}_2 = 0.408$ (0.05) [1.9 %]	$\hat{\alpha} = 0.587$ (0.02) [17.5 %] $\hat{a}_2 = 0.418$ (0.06) [4.5 %]
	$\alpha = 1$	$\hat{\alpha} = 1.005$ (0.04) [0.5 %] $\hat{a}_2 = 0.402$ (0.05) [0.5 %]	$\hat{\alpha} = 1.012$ (0.04) [1.2 %] $\hat{a}_2 = 0.403$ (0.05) [0.7 %]	$\hat{\alpha} = 1.021$ (0.04) [2.1 %] $\hat{a}_2 = 0.406$ (0.05) [1.6 %]	$\hat{\alpha} = 1.044$ (0.04) [4.4 %] $\hat{a}_2 = 0.412$ (0.05) [3.0 %]
	$\alpha = 2$	$\hat{\alpha} = 2.009$ (0.07) [0.5 %] $\hat{a}_2 = 0.401$ (0.05) [0.3 %]	$\hat{\alpha} = 2.006$ (0.07) [0.3 %] $\hat{a}_2 = 0.401$ (0.05) [0.3 %]	$\hat{\alpha} = 2.005$ (0.07) [0.3 %] $\hat{a}_2 = 0.402$ (0.05) [0.5 %]	$\hat{\alpha} = 1.997$ (0.07) [-0.2 %] $\hat{a}_2 = 0.401$ (0.05) [0.2 %]

Tabela 3.17: Simulação para durações Weibull, $\lambda = exp(0.4X_2)$ e 10 % de observações censuradas

A influência de censura fraca também não parece ter uma interpretação óbvia sobre a estimação quer de α quer de a_2 mas atendendo ao padrão observado para a variável dummy [12] é de notar que o enviesamento relativo na estimação de $\alpha = 0.5$, quando 50% das observações são *amontoadas*, passa de 12.4% sem censura para 17.5% na presença de 10% de observações censuradas, enquanto o enviesamento relativo do estimador de a_2 se reduz de 7% para 4.5%. Os resultados que se seguem reforçam esta interpretação.

[12] Que também é simétrica!

		Percentagem de heaping			
		0 %	10 %	20 %	50 %
$\lambda = exp(0.4X_2)$	$\alpha = 0.5$	$\hat{\alpha} = 0.502$	$\hat{\alpha} = 0.519$	$\hat{\alpha} = 0.538$	$\hat{\alpha} = 0.604$
		(0.02) [0.5 %]	(0.02) [3.9 %]	(0.02) [7.6 %]	(0.03) [20.7 %]
		$\hat{a}_2 = 0.403$	$\hat{a}_2 = 0.404$	$\hat{a}_2 = 0.406$	$\hat{a}_2 = 0.416$
		(0.06) [0.8 %]	(0.06) [0.9 %]	(0.06) [1.4 %]	(0.06) [4.0 %]
	$\alpha = 1$	$\hat{\alpha} = 1.004$	$\hat{\alpha} = 1.015$	$\hat{\alpha} = 1.026$	$\hat{\alpha} = 1.055$
		(0.04) [0.4 %]	(0.04) [1.5 %]	(0.04) [2.6 %]	(0.04) [5.5 %]
		$\hat{a}_2 = 0.403$	$\hat{a}_2 = 0.404$	$\hat{a}_2 = 0.405$	$\hat{a}_2 = 0.411$
		(0.06) [0.7 %]	(0.06) [1.0 %]	(0.06) [1.1 %]	(0.06) [2.8 %]
	$\alpha = 2$	$\hat{\alpha} = 2.012$	$\hat{\alpha} = 2.007$	$\hat{\alpha} = 2.006$	$\hat{\alpha} = 2.003$
		(0.08) [0.6 %]	(0.08) [0.4 %]	(0.08) [0.3 %]	(0.08) [0.1 %]
		$\hat{a}_2 = 0.403$	$\hat{a}_2 = 0.402$	$\hat{a}_2 = 0.403$	$\hat{a}_2 = 0.401$
		(0.06) [0.7 %]	(0.06) [0.6 %]	(0.06) [0.7 %]	(0.06) [0.3 %]

Tabela 3.18: Simulação para durações Weibull, $\lambda = exp(0.4X_2)$ e 30 % de observações censuradas

Esta tabela reforça o padrão já observado, censura, mesmo moderada, aumenta o enviesamento relativo do estimador de α e diminui o enviesamento relativo do estimador do parâmetro da variável explicativa simétrica. Repare-se nos valores estimados para 50% de observações *amontoadas*, enviesamentos relativos de 12.4% vs. 20.7%, na estimação de α e 7% vs. 4%, na estimação de a_2.

		Percentagem de heaping			
		0 %	10 %	20 %	50 %
$\lambda = exp(0.4X_2)$	$\alpha = 0.5$	$\hat{\alpha} = 0.502$	$\hat{\alpha} = 0.527$	$\hat{\alpha} = 0.557$	$\hat{\alpha} = 0.664$
		(0.03) [0.4 %]	(0.03) [5.5 %]	(0.03) [11.4 %]	(0.04) [32.8 %]
		$\hat{a}_2 = 0.404$	$\hat{a}_2 = 0.400$	$\hat{a}_2 = 0.401$	$\hat{a}_2 = 0.404$
		(0.07) [1.0 %]	(0.07) [0.1 %]	(0.07) [0.3 %]	(0.07) [1.1 %]
	$\alpha = 1$	$\hat{\alpha} = 1.004$	$\hat{\alpha} = 1.021$	$\hat{\alpha} = 1.035$	$\hat{\alpha} = 1.084$
		(0.05) [0.4 %]	(0.05) [2.1 %]	(0.05) [3.5 %]	(0.05) [8.4 %]
		$\hat{a}_2 = 0.403$	$\hat{a}_2 = 0.405$	$\hat{a}_2 = 0.405$	$\hat{a}_2 = 0.410$
		(0.07) [0.6 %]	(0.07) [1.1 %]	(0.07) [1.2 %]	(0.07) [2.6 %]
	$\alpha = 2$	$\hat{\alpha} = 2.009$	$\hat{\alpha} = 2.012$	$\hat{\alpha} = 2.013$	$\hat{\alpha} = 2.023$
		(0.10) [0.5 %]	(0.10) [0.6 %]	(0.10) [0.6 %]	(0.10) [1.2 %]
		$\hat{a}_2 = 0.404$	$\hat{a}_2 = 0.402$	$\hat{a}_2 = 0.401$	$\hat{a}_2 = 0.404$
		(0.07) [1.0 %]	(0.07) [0.6 %]	(0.07) [0.3 %]	(0.07) [1.0 %]

Tabela 3.19: Simulação para durações Weibull, $\lambda = exp(0.4X_2)$ e 50 % de observações censuradas

Censura severa vem reforçar de forma bem evidente o padrão observado em todos os casos anteriores, repare-se uma vez mais nos valores estimados para 50% de observações *amontoadas*, enviesamentos relativos de 12.4% vs. 32.8% na estimação de $\alpha = 0.5$ e 7% vs. 1.1% na estimação de a_2. Este padrão não é obviamente tão evidente nos casos de $\alpha = 1$ e $\alpha = 2$, uma vez que por um lado, os enviesamentos relativos são muito menos pronunciados e por outro lado, estimadores de MV são em geral enviesados e o desempenho destes estimadores é razoável ou mesmo bom no caso de parâmetros da dependências temporal elevados.

Verifiquemos se este padrão também é observado no caso da variável assimétrica (assimetria positiva) X_3.

3.5.3 Variável assimétrica como variável explicativa

Nesta subsecção, $\lambda = exp(0.4X_3)$. Os resultados das três tabelas que se seguem serão comparados com os da parte inferior das tabelas das páginas 30, 34 e 37 [13].

		Percentagem de heaping			
		0 %	10 %	20 %	50 %
$\lambda = exp(0.4X_3)$	$\alpha = 0.5$	$\hat{\alpha} = 0.502$ (0.02) [0.4 %] $\hat{a}_3 = 0.403$ (0.03) [0.7 %]	$\hat{\alpha} = 0.508$ (0.02) [1.6 %] $\hat{a}_3 = 0.346$ (0.05) [-13.4 %]	$\hat{\alpha} = 0.522$ (0.02) [4.5 %] $\hat{a}_3 = 0.308$ (0.05) [-23.0 %]	$\hat{\alpha} = 0.602$ (0.03) [20.4 %] $\hat{a}_3 = 0.256$ (0.04) [-35.9 %]
	$\alpha = 1$	$\hat{\alpha} = 1.005$ (0.04) [0.5 %] $\hat{a}_3 = 0.403$ (0.03) [0.6 %]	$\hat{\alpha} = 1.015$ (0.04) [1.5 %] $\hat{a}_3 = 0.384$ (0.04) [-3.9 %]	$\hat{\alpha} = 1.028$ (0.04) [2.8 %] $\hat{a}_3 = 0.369$ (0.04) [-7.9 %]	$\hat{\alpha} = 1.077$ (0.04) [7.7 %] $\hat{a}_3 = 0.339$ (0.05) [-15.3 %]
	$\alpha = 2$	$\hat{\alpha} = 2.009$ (0.07) [0.4 %] $\hat{a}_3 = 0.403$ (0.03) [0.7 %]	$\hat{\alpha} = 2.010$ (0.07) [0.5 %] $\hat{a}_3 = 0.401$ (0.03) [0.2 %]	$\hat{\alpha} = 2.012$ (0.07) [0.6 %] $\hat{a}_3 = 0.399$ (0.03) [-0.4 %]	$\hat{\alpha} = 2.016$ (0.07) [0.8 %] $\hat{a}_3 = 0.393$ (0.03) [-1.8 %]

Tabela 3.20: Simulação para durações Weibull, $\lambda = exp(0.4X_3)$ e 10 % de observações censuradas

Censura fraca não parece ter influência significativa sobre a estimação de a_3, os valores desta tabela reproduzem os das tabelas correspondentes ao estudo de dados não censurados. Nota-se contudo, uma pior performance do estimador de α, sobretudo no caso de 50% de observações *amontoadas*, o enviesamento relativo aumenta de 15.9% para 20.4%. O padrão do enviesamento relativo é o mesmo e os seus valores muito semelhantes.

[13] Variável X_3, *heaping* ignorado com $\alpha = 0.5$, $\alpha = 1$ e $\alpha = 2$, respectivamente.

		Percentagem de heaping			
		0 %	10 %	20 %	50 %
$\lambda = exp(0.4X_3)$	$\alpha = 0.5$	$\hat{\alpha} = 0.502$	$\hat{\alpha} = 0.510$	$\hat{\alpha} = 0.531$	$\hat{\alpha} = 0.659$
		(0.02) [0.4 %]	(0.03) [2.0 %]	(0.03) [6.2%]	(0.04) [31.7 %]
		$\hat{a_3} = 0.403$	$\hat{a_3} = 0.342$	$\hat{a_3} = 0.306$	$\hat{a_3} = 0.248$
		(0.03) [0.7 %]	(0.05) [-14.4 %]	(0.05) [-23.5%]	(0.04) [-37.9 %]
	$\alpha = 1$	$\hat{\alpha} = 1.003$	$\hat{\alpha} = 1.022$	$\hat{\alpha} = 1.041$	$\hat{\alpha} = 1.116$
		(0.04) [0.3 %]	(0.05) [2.2 %]	(0.05) [4.1 %]	(0.05) [11.6 %]
		$\hat{a_3} = 0.402$	$\hat{a_3} = 0.383$	$\hat{a_3} = 0.367$	$\hat{a_3} = 0.337$
		(0.03) [0.5 %]	(0.04) [-4.2 %]	(0.05) [-8.2 %]	(0.05) [-15.8 %]
	$\alpha = 2$	$\hat{\alpha} = 2.008$	$\hat{\alpha} = 2.015$	$\hat{\alpha} = 2.019$	$\hat{\alpha} = 2.028$
		(0.08) [0.4 %]	(0.08) [0.7 %]	(0.08) [1.0 %]	(0.08) [1.4 %]
		$\hat{a_3} = 0.403$	$\hat{a_3} = 0.401$	$\hat{a_3} = 0.400$	$\hat{a_3} = 0.393$
		(0.03) [0.8 %]	(0.03) [0.4 %]	(0.03) [0.0 %]	(0.03) [-1.7 %]

Tabela 3.21: Simulação para durações Weibull, $\lambda = exp(0.4X_3)$ e 30 % de observações censuradas

Na estimação de a_3 censura moderada também não conduz a resultados significativamente diferentes dos obtidos para dados não censurados. No caso da estimação de α pelo contrário, com 50% de *heaping* , o enviesamento relativo aumenta, em valor absoluto; se $\alpha = 1$, aumenta de 6.5% para 11.6% e se $\alpha = 0.5$, aumenta de 15.9% para 31.7%.

Note-se que censura moderada combinada com *heaping* mesmo que moderado e padrão de dependência temporal negativo conduziu em 2.6% dos casos à não convergência dos algoritmos implementados neste estudo.

		Percentagem de heaping			
		0 %	10 %	20 %	50 %
$\lambda = exp(0.4X_3)$	$\alpha = 0.5$	$\hat{\alpha} = 0.503$ (0.03) [0.6 %]	$\hat{\alpha} = 0.507$ (0.03) [1.5 %]	$\hat{\alpha} = 0.534$ (0.03) [6.7%]	$\hat{\alpha} = 0.729$ (0.06) [45.8 %]
		$\hat{a}_3 = 0.403$ (0.03) [0.8 %]	$\hat{a}_3 = 0.340$ (0.06) [-15.0 %]	$\hat{a}_3 = 0.298$ (0.05) [-25.6 %]	$\hat{a}_3 = 0.23$ (0.04) [-42.5 %]
	$\alpha = 1$	$\hat{\alpha} = 1.006$ (0.05) [0.6 %]	$\hat{\alpha} = 1.030$ (0.06) [3.0 %]	$\hat{\alpha} = 1.062$ (0.06) [6.2 %]	$\hat{\alpha} = 1.185$ (0.06) [18.5 %]
		$\hat{a}_3 = 0.403$ (0.03) [0.8 %]	$\hat{a}_3 = 0.382$ (0.04) [-4.6 %]	$\hat{a}_3 = 0.364$ (0.05) [-8.9 %]	$\hat{a}_3 = 0.333$ (0.05) [-16.8 %]
	$\alpha = 2$	$\hat{\alpha} = 2.009$ (0.10) [0.5 %]	$\hat{\alpha} = 2.019$ (0.10) [0.9 %]	$\hat{\alpha} = 2.030$ (0.10) [1.5 %]	$\hat{\alpha} = 2.060$ (0.10) [3.0 %]
		$\hat{a}_3 = 0.403$ (0.03) [0.7 %]	$\hat{a}_3 = 0.402$ (0.03) [0.5 %]	$\hat{a}_3 = 0.400$ (0.03) [0.1 %]	$\hat{a}_3 = 0.394$ (0.04) [-1.4 %]

Tabela 3.22: Simulação para durações Weibull, $\lambda = exp(0.4X_3)$ e 50 % de observações censuradas

Censura severa parece confirmar o padrão observado até aqui para o caso desta variável, a saber:

- Pouca influência da censura no enviesamento relativo do estimador de a_3. Ainda assim o enviesamento relativo cresce ligeiramente com o aumento do número de observações censuradas.

- Aumenta o enviesamento relativo do estimador de α. No caso mais extremado (50% de *heaping*) o enviesamento relativo aumenta de 15.9% para 45.8%.

- *Heaping* é o factor dominante para censura fraca e moderada.

- Problemas de convergência quando associada a *heaping* moderado ou severo. Tivemos até 7.1% de simulações em que os algoritmos não convergiram.

Capítulo 4

Conclusões

No contexto das aplicações empíricas de modelos de duração é frequente os investigadores fazerem pressupostos simplificadores, como por exemplo ignorarem a existência de efeito de *heaping* nos dados. Neste estudo, pretendeu-se, através de um conjunto de simulações de Monte Carlo, responder a uma questão pouco estudada na literatura:

A partir de que quantidade de heaping é de facto desaconselhável ignorar o heaping ?

Estudámos o desempenho dos estimadores de máxima verosimilhança, MV, para percentagens de *heaping* crescentes, 0%,10%, 20%, 30%, 40% e 50%, para um modelo de *hazard* proporcional, Weibull, com três tipos de dependência temporal: negativa , ausência de dependência e dependência temporal positiva. Para variáveis explicativas escolhemos X_1, variável dummy com 50% de $1's$, X_2 com distribuição normal *standard* e X_3, com uma distribuição qui-quadrado com dois graus de liberdade. Em todos os casos os resultados diferem entre as variáveis com distribuições simétricas, (X_1 e X_2) e X_3 que tem uma distribuição assimétrica positiva. Mais *heaping* leva sempre a enviesamentos crescentes, em valor absoluto, nos estimadores, no caso assimétrico de uma forma sempre mais pronunciada do que do caso simétrico. Quanto ao efeito da escolha de diferentes dependências temporais, o modelo é tanto mais insensível ao *heaping* quanto maior for α [1], i. é, o *heaping* enviesa tanto mais os estimadores quanto mais negativa for a dependência temporal. No caso de $\alpha = 0.5$, dependencia negativa, o enviesamento é positivo na estimação do parâmetro temporal e dos coeficiente da variáveis explicativa simétricas mas negativo no caso da variável assimétrica. O caso $\alpha = 1$, durações com distribuição exponencial, é semelhante ao de $\alpha = 0.5$ mas os enviesamentos são bastante menores (em termos absolutos). No caso exponencial se os regressores seguirem distribuições simétricas o *heaping* não parece afectar o enviesamento dos estimadores em estudo, o que está em linha com os resultados deduzidos por Augustin e Wolff (2003). No caso de $\alpha = 2$, dependência positiva, não se pode concluir que enviesamento dos estimadores dos parâmetros do modelo variam em função de quantidades de *heaping* crescentes. Com o padrão de *heaping* escolhido neste estudo este praticamente não interfere com a estimação, pode até dizer-se que, nalguns casos "melhora" a qualidade dos estimadores, reduzindo-lhes o enviesamento! Mais uma vez, respondendo à questão formulada originalmente: Ignorar o *heaping* não conduz forçosamente a uma degradação da performance dos estimadores. A performance está sobretudo dependente da dimensão de α, quanto maior for, menor será o impacto do *heaping* e tipo de distribuição dos regressores do modelo.

Quanto à esperada inconsistência dos estimadores que ignoram o *heaping* (*naïves*, na literatura) encontrámos evidência empírica dessa inconsistência mesmo com *heaping* simétrico e para percentagens de *heaping* pequenas; em linha com os resultados obtidos por Augustin e Wolff (2003), com particular realce para a forte evidência no caso de dependência temporal negativa em que o enviesamento estimado do estimador do parâmetro da variável assimétrica aumenta com o tamanho da amostra em vez de diminuir.

Finalmente estudámos o impacto da censura à direita na performance dos estimadores de MV e comparámos esses resultados com os obtidos para dados não censurados. Como era de esperar, os resultados diferem para dois casos considerados: variáveis simétricas vs. variáveis assimétricas. No caso simétrico, percentagens crescentes de censura fazem aumentar o enviesamento do estimador de α (relativamente à situação

[1] Parâmetro temporal da distribuição Weibull

de ausência de censura) mas o enviesamento do estimador do coeficiente da variável *dummy* diminui. A presença de censura pode representar um ganho de qualidade (relativa) na estimação dos parâmetros associados aos regressores com distribuições simétricas. No caso da variável explicativa assimétrica, censura crescente leva ao aumenta do enviesamento do estimador de α e praticamente não interfere com a estimação do parâmetro do regressor; a presença de censura não tem qualquer tipo de vantagem para um modelo com este tipo de regressores assimétricos .

Para futuras linhas de investigação sugerimos:
A generalização dos resultados a outros modelos, em particular a uma generalização natural do modelo Weibull, a Gamma generalizada, que esboçámos mas cujos resultados não apresentamos neste trabalho por não termos superado, nalguns casos, problemas de convergência dos algoritmos implementados.
O estudo das propriedades dos estimadores que generalizam os que estudámos e modelam o *heaping* .
A melhoria do enquadramento teórico para que os resultados recebam um suporte conveniente e não passem de mais o resultado de um estudo por simulação com pouco valor na modelação de dados *amontoados.*
Aos estatísticos que procuram modelar **dados reais** afectados por *heaping* , após terem deduzido de alguma forma o seu padrão, a facilidade de implementação dos métodos descritos encoraja a um estudo prévio, por simulação de Monte Carlo, antes de implementar um modelo que ignore o *heaping* .
Para concluir, uma nota de agradecimento aos mentores e colaboradores do *The R Project for Statistical Computing*, software livre, intuitivo e versátil que nos poupou à árdua e entediante tarefa de programar os algoritmos de estimação.

Apêndice A

Heaping ignorado

Segue-se os resultados **completos** da estimação para o parâmetro temporal $\alpha = 0.5$, de seguida $\alpha = 1$ e finalmente $\alpha = 2$. Entre parênteses curvos uma estimativa do desvio padrão e entre parênteses rectos uma estimativa do enviesamento relativo:

A.1 $\alpha = 0.5$

Percentagem de *heaping*

	0%	5%	10%	20%	30%	40%	50%
X_1	$\hat{\alpha} = 0.502$	$\hat{\alpha} = 0.508$	$\hat{\alpha} = 0.514$	$\hat{\alpha} = 0.526$	$\hat{\alpha} = 0.538$	$\hat{\alpha} = 0.551$	$\hat{\alpha} = 0.564$
	(0.017)	(0.018)	(0.018)	(0.018)	(0.018)	(0.019)	(0.019)
	[0.4%]	[1.6%]	[2.8%]	[5.1%]	[7.7%]	[10.2%]	[12.9%]
	$\hat{a}_1 = 0.202$	$\hat{a}_1 = 0.204$	$\hat{a}_1 = 0.205$	$\hat{a}_1 = 0.206$	$\hat{a}_1 = 0.208$	$\hat{a}_1 = 0.213$	$\hat{a}_1 = 0.216$
	(0.092)	(0.092)	(0.091)	(0.092)	(0.096)	(0.096)	(0.098)
	[1%]	[1.8%]	[2.4%]	[3.1%]	[3.9%]	[6.4%]	[7.9%]
	$\alpha = 0.5$ - $a_1 = 0.2$ / $\alpha = 0.5$ - $a_1 = 0.4$						
	$\hat{\alpha} = 0.502$	$\hat{\alpha} = 0.508$	$\hat{\alpha} = 0.515$	$\hat{\alpha} = 0.529$	$\hat{\alpha} = 0.542$	$\hat{\alpha} = 0.556$	$\hat{\alpha} = 0.572$
	(0.018)	(0.018)	(0.018)	(0.018)	(0.019)	(0.019)	(0.019)
	[0.5%]	[1.7%]	[3%]	[5.7%]	[8.5%]	[11.3%]	[14.4%]
	$\hat{a}_1 = 0.4$	$\hat{a}_1 = 0.405$	$\hat{a}_1 = 0.409$	$\hat{a}_1 = 0.415$	$\hat{a}_1 = 0.419$	$\hat{a}_1 = 0.427$	$\hat{a}_1 = 0.434$
	(0.092)	(0.092)	(0.092)	(0.094)	(0.097)	(0.097)	(0.099)
	[0.1%]	[1.4%]	[2.2%]	[3.8%]	[4.7%]	[6.7%]	[8.5%]

Tabela A.1: Simulação para durações Weibull, $\alpha = 0.5$ - **parte superior** $\lambda = exp(0.2X_1)$ - **parte inferior** $\lambda = exp(0.4X_1)$.

Percentagem de *heaping*

	0%	5%	10%	20%	30%	40%	50%
X_2	$\hat{\alpha} = 0.502$	$\hat{\alpha} = 0.508$	$\hat{\alpha} = 0.513$	$\hat{\alpha} = 0.524$	$\hat{\alpha} = 0.535$	$\hat{\alpha} = 0.547$	$\hat{\alpha} = 0.559$
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.019)	(0.019)
	[0.4%]	[1.5%]	[2.6%]	[4.7%]	[7.1%]	[9.4%]	[11.9%]
	$\hat{a}_2 = 0.201$	$\hat{a}_2 = 0.202$	$\hat{a}_2 = 0.203$	$\hat{a}_2 = 0.206$	$\hat{a}_2 = 0.209$	$\hat{a}_2 = 0.211$	$\hat{a}_2 = 0.215$
	(0.045)	(0.048)	(0.047)	(0.047)	(0.048)	(0.048)	(0.049)
	[0.7%]	[0.8%]	[1.6%]	[2.8%]	[4.4%]	[5.6%]	[7.4%]
	$\alpha = 0.5$ – $a_2 = 0.2$ / $\alpha = 0.5$ – $a_2 = 0.4$						
	$\hat{\alpha} = 0.502$	$\hat{\alpha} = 0.508$	$\hat{\alpha} = 0.514$	$\hat{\alpha} = 0.525$	$\hat{\alpha} = 0.537$	$\hat{\alpha} = 0.549$	$\hat{\alpha} = 0.562$
	(0.018)	(0.018)	(0.018)	(0.018)	(0.019)	(0.018)	(0.019)
	[0.4%]	[1.5%]	[2.7%]	[4.9%]	[7.4%]	[9.8%]	[12.4%]
	$\hat{a}_2 = 0.402$	$\hat{a}_2 = 0.404$	$\hat{a}_2 = 0.406$	$\hat{a}_2 = 0.411$	$\hat{a}_2 = 0.416$	$\hat{a}_2 = 0.421$	$\hat{a}_2 = 0.428$
	(0.047)	(0.047)	(0.048)	(0.048)	(0.048)	(0.05)	(0.051)
	[0.6%]	[1%]	[1.4%]	[2.7%]	[4%]	[5.2%]	[7%]

Tabela A.2: Simulação para durações Weibull, $\alpha = 0.5$ - **parte superior** $\lambda = exp(0.2X_2)$ - **parte inferior** $\lambda = exp(0.4X_2)$.

Percentagem de *heaping*

	0%	5%	10%	20%	30%	40%	50%
X_3	$\hat{\alpha} = 0.502$	$\hat{\alpha} = 0.509$	$\hat{\alpha} = 0.517$	$\hat{\alpha} = 0.532$	$\hat{\alpha} = 0.549$	$\hat{\alpha} = 0.566$	$\hat{\alpha} = 0.585$
	(0.017)	(0.018)	(0.018)	(0.019)	(0.019)	(0.02)	(0.021)
	[0.3%]	[1.8%]	[3.3%]	[6.5%]	[9.7%]	[13.2%]	[17.1%]
	$\hat{a}_3 = 0.201$	$\hat{a}_3 = 0.2$	$\hat{a}_3 = 0.197$	$\hat{a}_3 = 0.195$	$\hat{a}_3 = 0.193$	$\hat{a}_3 = 0.191$	$\hat{a}_3 = 0.191$
	(0.024)	(0.024)	(0.024)	(0.024)	(0.025)	(0.024)	(0.024)
	[0.7%]	[0%]	[−1.3%]	[−2.7%]	[−3.6%]	[−4.6%]	[−4.7%]
				$\alpha = 0.5$ – $a_3 = 0.2$ / $\alpha = 0.5$ – $a_3 = 0.4$			
	$\hat{\alpha} = 0.502$	$\hat{\alpha} = 0.503$	$\hat{\alpha} = 0.506$	$\hat{\alpha} = 0.518$	$\hat{\alpha} = 0.535$	$\hat{\alpha} = 0.556$	$\hat{\alpha} = 0.580$
	(0.018)	(0.02)	(0.02)	(0.021)	(0.021)	(0.022)	(0.022)
	[0.4%]	[0.6%]	[1.1%]	[3.6%]	[6.9%]	[11.1%]	[15.9%]
	$\hat{a}_3 = 0.402$	$\hat{a}_3 = 0.372$	$\hat{a}_3 = 0.346$	$\hat{a}_3 = 0.31$	$\hat{a}_3 = 0.285$	$\hat{a}_3 = 0.268$	$\hat{a}_3 = 0.256$
	(0.026)	(0.046)	(0.052)	(0.052)	(0.048)	(0.044)	(0.04)
	[0.5%]	[−7%]	[−13.4%]	[−22.6%]	[−28.8%]	[−32.9%]	[−36%]

Tabela A.3: Simulação para durações Weibull, $\alpha = 0.5$ - **parte superior** $\lambda = exp(0.2X_3)$ - **parte inferior** $\lambda = exp(0.4X_3)$.

A.2 $\alpha = 1$ (Exponencial)

Percentagem de *heaping*

X_1	0%	5%	10%	20%	30%	40%	50%
	$\hat{\alpha} = 1.004$	$\hat{\alpha} = 1.007$	$\hat{\alpha} = 1.01$	$\hat{\alpha} = 1.017$	$\hat{\alpha} = 1.025$	$\hat{\alpha} = 1.031$	$\hat{\alpha} = 1.038$
	(0.035)	(0.034)	(0.035)	(0.035)	(0.034)	(0.035)	(0.034)
	[0.4%]	[0.7%]	[1%]	[1.7%]	[2.5%]	[3.1%]	[3.8%]
	$\hat{a}_1 = 0.201$	$\hat{a}_1 = 0.2$	$\hat{a}_1 = 0.202$	$\hat{a}_1 = 0.204$	$\hat{a}_1 = 0.204$	$\hat{a}_1 = 0.205$	$\hat{a}_1 = 0.207$
	(0.091)	(0.09)	(0.092)	(0.091)	(0.092)	(0.093)	(0.094)
	[0.7%]	[0.1%]	[1.2%]	[2.2%]	[2.1%]	[2.4%]	[3.7%]
	$\alpha = 1$ – $a_1 = 0.2$						
	$\alpha = 1$ – $a_1 = 0.4$						
	$\hat{\alpha} = 1.004$	$\hat{\alpha} = 1.008$	$\hat{\alpha} = 1.012$	$\hat{\alpha} = 1.02$	$\hat{\alpha} = 1.027$	$\hat{\alpha} = 1.036$	$\hat{\alpha} = 1.043$
	(0.035)	(0.035)	(0.035)	(0.035)	(0.034)	(0.034)	(0.034)
	[0.4%]	[0.8%]	[1.2%]	[2%]	[2.7%]	[3.6%]	[4.3%]
	$\hat{a}_1 = 0.4$	$\hat{a}_1 = 0.404$	$\hat{a}_1 = 0.404$	$\hat{a}_1 = 0.407$	$\hat{a}_1 = 0.408$	$\hat{a}_1 = 0.412$	$\hat{a}_1 = 0.415$
	(0.092)	(0.091)	(0.093)	(0.093)	(0.093)	(0.092)	(0.096)
	[0.1%]	[0.9%]	[1%]	[1.7%]	[2%]	[3.1%]	[3.8%]

Tabela A.4: Simulação para durações Weibull (exponenciais), $\alpha = 1$ - **parte superior** $\lambda = exp(0.2X_1)$ - **parte inferior** $\lambda = exp(0.4X_1)$.

	Percentagem de *heaping*						
	0%	5%	10%	20%	30%	40%	50%
X_2	$\hat{\alpha} = 1.003$	$\hat{\alpha} = 1.007$	$\hat{\alpha} = 1.01$	$\hat{\alpha} = 1.016$	$\hat{\alpha} = 1.022$	$\hat{\alpha} = 1.029$	$\hat{\alpha} = 1.035$
	(0.035)	(0.035)	(0.035)	(0.034)	(0.034)	(0.035)	(0.035)
	[0.3%]	[0.7%]	[1%]	[1.6%]	[2.2%]	[2.9%]	[3.5%]
	$\hat{a}_2 = 0.202$	$\hat{a}_2 = 0.202$	$\hat{a}_2 = 0.204$	$\hat{a}_2 = 0.203$	$\hat{a}_2 = 0.204$	$\hat{a}_2 = 0.205$	$\hat{a}_2 = 0.206$
	(0.046)	(0.047)	(0.047)	(0.045)	(0.045)	(0.046)	(0.047)
	[0.9%]	[1.1%]	[1.8%]	[1.3%]	[2.2%]	[2.3%]	[2.8%]
	$\alpha = 1$ – $a_2 = 0.2$						
	$\alpha = 1$ – $a_2 = 0.4$						
	$\hat{\alpha} = 1.005$	$\hat{\alpha} = 1.007$	$\hat{\alpha} = 1.01$	$\hat{\alpha} = 1.017$	$\hat{\alpha} = 1.024$	$\hat{\alpha} = 1.031$	$\hat{\alpha} = 1.037$
	(0.035)	(0.035)	(0.036)	(0.035)	(0.035)	(0.035)	(0.034)
	[0.5%]	[0.7%]	[1%]	[1.7%]	[2.4%]	[3.1%]	[3.7%]
	$\hat{a}_2 = 0.402$	$\hat{a}_2 = 0.403$	$\hat{a}_2 = 0.404$	$\hat{a}_2 = 0.406$	$\hat{a}_2 = 0.407$	$\hat{a}_2 = 0.410$	$\hat{a}_2 = 0.412$
	(0.047)	(0.047)	(0.048)	(0.046)	(0.048)	(0.049)	(0.049)
	[0.4%]	[0.6%]	[1%]	[1.5%]	[1.9%]	[2.4%]	[3.1%]

Tabela A.5: Simulação para durações Weibull (exponenciais), $\alpha = 1$ - **parte superior** $\lambda = exp(0.2X_2)$ - **parte inferior** $\lambda = exp(0.4X_2)$.

	Percentagem de *heaping*						
	0%	5%	10%	20%	30%	40%	50%
X_3	$\hat{\alpha} = 1.004$	$\hat{\alpha} = 1.009$	$\hat{\alpha} = 1.014$	$\hat{\alpha} = 1.025$	$\hat{\alpha} = 1.034$	$\hat{\alpha} = 1.045$	$\hat{\alpha} = 1.055$
	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)
	[0.4%]	[0.9%]	[1.4%]	[2.5%]	[3.4%]	[4.5%]	[5.5%]
	$\hat{a}_3 = 0.202$	$\hat{a}_3 = 0.202$	$\hat{a}_3 = 0.202$	$\hat{a}_3 = 0.203$	$\hat{a}_3 = 0.202$	$\hat{a}_3 = 0.204$	$\hat{a}_3 = 0.204$
	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)
	[0.8%]	[1%]	[0.9%]	[1.3%]	[1.2%]	[1.9%]	[2%]
	$\alpha = 1$ – $a_3 = 0.2$						
	$\alpha = 1$ – $a_3 = 0.4$						
	$\hat{\alpha} = 1.005$	$\hat{\alpha} = 1.009$	$\hat{\alpha} = 1.014$	$\hat{\alpha} = 1.024$	$\hat{\alpha} = 1.036$	$\hat{\alpha} = 1.05$	$\hat{\alpha} = 1.065$
	(0.035)	(0.037)	(0.036)	(0.037)	(0.038)	(0.038)	(0.037)
	[0.5%]	[0.9%]	[1.4%]	[2.4%]	[3.6%]	[5%]	[6.5%]
	$\hat{a}_3 = 0.402$	$\hat{a}_3 = 0.392$	$\hat{a}_3 = 0.386$	$\hat{a}_3 = 0.370$	$\hat{a}_3 = 0.356$	$\hat{a}_3 = 0.347$	$\hat{a}_3 = 0.338$
	(0.027)	(0.035)	(0.038)	(0.043)	(0.044)	(0.045)	(0.044)
	[0.6%]	[−1.9%]	[−3.6%]	[−7.6%]	[−10.9%]	[−13.2%]	[−15.4%]

Tabela A.6: Simulação para durações Weibull (exponenciais), $\alpha = 1$ - **parte superior** $\lambda = exp(0.2X_3)$ - **parte inferior** $\lambda = exp(0.4X_3)$.

A.3 $\alpha = 2$

Percentagem de *heaping*

	0%	5%	10%	20%	30%	40%	50%
X_1	$\hat{\alpha} = 2.007$	$\hat{\alpha} = 2.006$	$\hat{\alpha} = 2.004$	$\hat{\alpha} = 2.002$	$\hat{\alpha} = 1.999$	$\hat{\alpha} = 1.996$	$\hat{\alpha} = 1.993$
	(0.07)	(0.07)	(0.069)	(0.069)	(0.07)	(0.07)	(0.068)
	[0.3%]	[0.3%]	[0.2%]	[0.1%]	[−0.1%]	[−0.2%]	[−0.3%]
	$\hat{a}_1 = 0.2$	$\hat{a}_1 = 0.2$	$\hat{a}_1 = 0.2$	$\hat{a}_1 = 0.199$	$\hat{a}_1 = 0.203$	$\hat{a}_1 = 0.203$	$\hat{a}_1 = 0.2$
	(0.091)	(0.091)	(0.09)	(0.091)	(0.091)	(0.09)	(0.091)
	[0.1%]	[0.2%]	[0.2%]	[−0.5%]	[1.4%]	[1.3%]	[0.1%]
				$\alpha = 2$ – $a_1 = 0.2$			
				$\alpha = 2$ – $a_1 = 0.4$			
	$\hat{\alpha} = 2.009$	$\hat{\alpha} = 2.005$	$\hat{\alpha} = 2.006$	$\hat{\alpha} = 2.003$	$\hat{\alpha} = 2.002$	$\hat{\alpha} = 1.999$	$\hat{\alpha} = 1.996$
	(0.07)	(0.07)	(0.07)	(0.069)	(0.069)	(0.07)	(0.069)
	[0.5%]	[0.3%]	[0.3%]	[0.2%]	[0.1%]	[0%]	[−0.2%]
	$\hat{a}_1 = 0.403$	$\hat{a}_1 = 0.399$	$\hat{a}_1 = 0.404$	$\hat{a}_1 = 0.401$	$\hat{a}_1 = 0.401$	$\hat{a}_1 = 0.401$	$\hat{a}_1 = 0.404$
	(0.092)	(0.092)	(0.092)	(0.092)	(0.092)	(0.095)	(0.093)
	[0.8%]	[−0.3%]	[1%]	[0.2%]	[0.3%]	[0.2%]	[1%]

Tabela A.7: Simulação para durações Weibull, $\alpha = 2$ - **parte superior** $\lambda = exp(0.2X_1)$ - **parte inferior** $\lambda = exp(0.4X_1)$.

Percentagem de *heaping*

	0%	5%	10%	20%	30%	40%	50%
X_2	$\hat{\alpha} = 2.008$	$\hat{\alpha} = 2.008$	$\hat{\alpha} = 2.005$	$\hat{\alpha} = 2.002$	$\hat{\alpha} = 1.996$	$\hat{\alpha} = 1.995$	$\hat{\alpha} = 1.991$
	(0.07)	(0.071)	(0.071)	(0.069)	(0.068)	(0.07)	(0.068)
	[0.4%]	[0.4%]	[0.2%]	[0.1%]	[−0.2%]	[−0.3%]	[−0.4%]
	$\hat{a}_2 = 0.201$	$\hat{a}_2 = 0.201$	$\hat{a}_2 = 0.2$	$\hat{a}_2 = 0.201$	$\hat{a}_2 = 0.2$	$\hat{a}_2 = 0.201$	$\hat{a}_2 = 0.201$
	(0.045)	(0.045)	(0.047)	(0.045)	(0.045)	(0.046)	(0.046)
	[0.7%]	[0.7%]	[0.2%]	[0.3%]	[0.2%]	[0.5%]	[0.6%]
	$\alpha = 2$ – $a_2 = 0.2$ / $\alpha = 2$ – $a_2 = 0.4$						
	$\hat{\alpha} = 2.006$	$\hat{\alpha} = 2.006$	$\hat{\alpha} = 2.006$	$\hat{\alpha} = 2.004$	$\hat{\alpha} = 1.999$	$\hat{\alpha} = 1.997$	$\hat{\alpha} = 1.993$
	(0.07)	(0.07)	(0.071)	(0.07)	(0.069)	(0.069)	(0.068)
	[0.3%]	[0.3%]	[0.3%]	[0.2%]	[0%]	[−0.2%]	[−0.4%]
	$\hat{a}_2 = 0.401$	$\hat{a}_2 = 0.4$	$\hat{a}_2 = 0.401$	$\hat{a}_2 = 0.402$	$\hat{a}_2 = 0.401$	$\hat{a}_2 = 0.401$	$\hat{a}_2 = 0.399$
	(0.047)	(0.048)	(0.048)	(0.048)	(0.048)	(0.047)	(0.048)
	[0.4%]	[0%]	[0.4%]	[0.5%]	[0.4%]	[0.3%]	[−0.2%]

Tabela A.8: Simulação para durações Weibull, $\alpha = 2$ - **parte superior** $\lambda = exp(0.2X_2)$ - **parte inferior** $\lambda = exp(0.4X_2)$.

Percentagem de *heaping*

	0%	5%	10%	20%	30%	40%	50%
X_3	$\hat{\alpha} = 2.008$	$\hat{\alpha} = 2.008$	$\hat{\alpha} = 2.006$	$\hat{\alpha} = 2.005$	$\hat{\alpha} = 2.005$	$\hat{\alpha} = 2.003$	$\hat{\alpha} = 2.003$
	(0.071)	(0.07)	(0.071)	(0.07)	(0.07)	(0.069)	(0.071)
	[0.4%]	[0.4%]	[0.3%]	[0.3%]	[0.2%]	[0.1%]	[0.1%]
	$\hat{a}_3 = 0.202$	$\hat{a}_3 = 0.202$	$\hat{a}_3 = 0.202$	$\hat{a}_3 = 0.202$	$\hat{a}_3 = 0.202$	$\hat{a}_3 = 0.201$	$\hat{a}_3 = 0.201$
	(0.024)	(0.025)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)
	[0.8%]	[1.1%]	[0.8%]	[0.8%]	[0.8%]	[0.7%]	[0.5%]
				$\alpha = 2$ – $a_3 = 0.2$			
				$\alpha = 2$ – $a_3 = 0.4$			
	$\hat{\alpha} = 2.009$	$\hat{\alpha} = 2.008$	$\hat{\alpha} = 2.008$	$\hat{\alpha} = 2.009$	$\hat{\alpha} = 2.011$	$\hat{\alpha} = 2.012$	$\hat{\alpha} = 2.012$
	(0.071)	(0.069)	(0.069)	(0.071)	(0.069)	(0.071)	(0.071)
	[0.4%]	[0.4%]	[0.4%]	[0.5%]	[0.5%]	[0.6%]	[0.6%]
	$\hat{a}_3 = 0.403$	$\hat{a}_3 = 0.401$	$\hat{a}_3 = 0.401$	$\hat{a}_3 = 0.398$	$\hat{a}_3 = 0.396$	$\hat{a}_3 = 0.395$	$\hat{a}_3 = 0.393$
	(0.027)	(0.027)	(0.028)	(0.029)	(0.029)	(0.03)	(0.03)
	[0.7%]	[0.4%]	[0.1%]	[−0.4%]	[−1%]	[−1.3%]	[−1.8%]

Tabela A.9: Simulação para durações Weibull, $\alpha = 2$ - **parte superior** $\lambda = exp(0.2X_3)$ - **parte inferior** $\lambda = exp(0.4X_3)$.

Referências

Augustin, T., & Schneeweiß, H. (2005). *Some recent advances in measurement error models and methods.* http://www.stat.uni-muenchen.de/sfb386/papers/dsp/paper452.pdf.

Augustin, T., & Wolff, J. (2003). Heaping and its consequences for duration analysis. a simulation study. *Allgemeines statistisches Archiv*, *87*(1), 59-86.

Battistin, E., Miniaci, R., & Weber, G. (2003). *What do we learn from recall consumption data?* (Temi di discussione (Economic working papers)). Bank of Italy, Economic Research Department. http://ideas.repec.org/p/bdi/wptemi/td_466_03.html,.

Brewer, D. D., & Roberts, J. M. (2001). Measures and tests of heaping in discrete quantitative distributions. *Journal of Applied Statistics*, *28*(7), 887-896.

Gill, R., Petoussis, K., & Zeelenberg, C. (2004). *Statistical analysis of heaped duration data.* http://citeseer.ist.psu.edu/petoussis97statistical.html.

Gleitman, H. (1993). *Psicologia.* Lisboa: Fundação Calouste Gulbenkian.

Heitjan, D. F., & Rubin, D. B. (1990). Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association*, *85*(410), 304-314.

Heitjan, D. F., & Rubin, D. B. (1991). Ignorability and coarse data. *The Annals of Statistics*, *19*(4), 2244-2253.

Horowitz, J. L., & Neumann, G. R. (1989). Specification testing in censored regression models: Parametric and semiparametric methods. *Journal of Applied Econometrics*, *4*, S61-S86.

Jürges, H. (2004). *Objects in the mirror are closer than they appear: Unemployment, retrospective error, and life satisfaction.* (forthcoming in the Journal of the Royal Statistical Society)

Kiefer, N. M. (1988). Economic duration data and hazard functions. *Journal of Economic Literature*, *26*(2), 646-79.

Kraus, F., & Steiner, V. (1998). Modelling heaping effects in unemployment duration models - with an application to retrospective event data in the german socio-economic panel. *Jahrbücher für Nationalökonomie und Statistik*, *217*(5), 550-573.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data, second edition.* New York, NY, USA: John Wiley & Sons, Inc.

Ramalho, E. A., & Smith, R. (2003). *Discrete choice non-response* (CeMMAP working papers No. CWP07/03). Centre for Microdata Methods and Practice, Institute for Fiscal Studies. (available at http://ideas.repec.org/p/ifs/cemmap/07-03.html)

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

Ryu, H. K., & Slottje, D. J. (2000). Estimating the density of unemployment duration based on contaminated samples or small samples. *Journal of Econometrics*, *95*(1), 131-156.

Torelli, N., & Trivellato, U. (1993). Modelling inaccuracies in job-search duration data. *Journal of Econometrics*, *59*(1-2), 187-211.